

# Scientific Computing for Biologists

## Data as Vectors

Instructor: Paul M. Magwene

07 September 2010

# Overview of Lecture

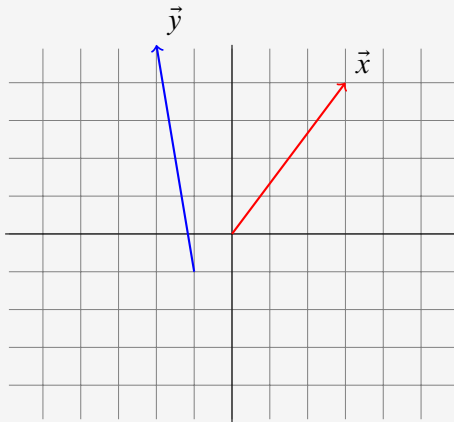
- Vector Geometry
  - Vectors are directed line segments
  - Vector length
- Vector Arithmetic
  - Addition, subtraction
  - Scalar multiplication
  - Linear combinations of vectors
  - Dot product and projection
- Vector representations of multivariate data
  - Variable space/Subject space representations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

# Hands-on Session

- Vector operations in R and Python
- Writing functions in R and Python
- Visualizing univariate distributions on R
- Linear regression and t-tests in R

# Vector Geometry

Vectors are directed line segments.

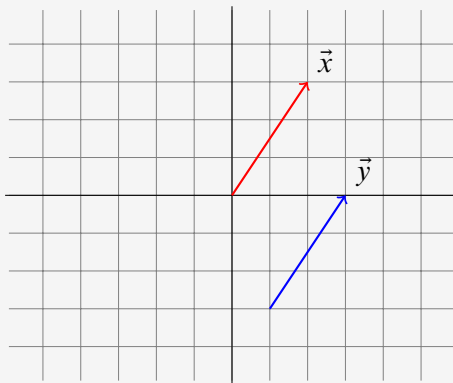


All of the figures and algebraic formulas I show you apply to  $n$ -dimensional vectors.

# Vector Geometry

Vectors have direction and length:

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

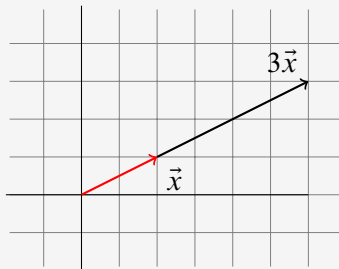


Often starting point is ignored, in which case  $\vec{x} = \vec{y}$ .

# Scalar Multiplication of a Vector

Let  $k$  be a scalar.

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

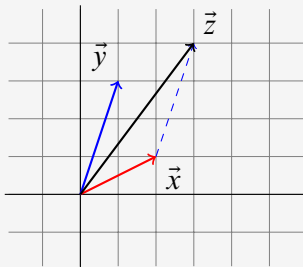


$$\vec{x} = [2, 1]'; \quad 3\vec{x} = [6, 3]'$$

## Vector Addition

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

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

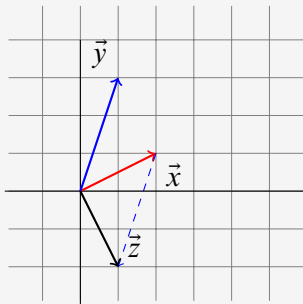


Addition follows the 'head-to-tail' rule.

# Vector Subtraction

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

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



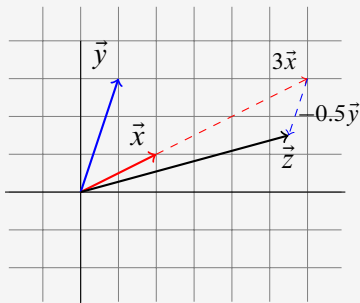
Follow the addition rule for  $-1\vec{y}$ .



# Linear Combinations of Vectors

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

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

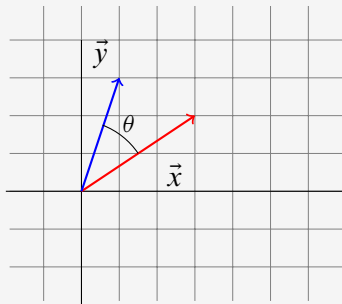


# Dot Product

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

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

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



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

# Useful Geometric Quantities as Dot Product

Length:

$$|\vec{x}|^2 = \vec{x} \cdot \vec{x} = x_1^2 + x_2^2 + \cdots + x_n^2$$

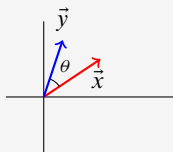
$$|\vec{y}|^2 = \vec{y} \cdot \vec{y}$$

Distance:

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

Angle:

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



# Dot Product Properties

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

$$\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x} \quad (\text{commutative})$$

$$\vec{x} \cdot (\vec{y} + \vec{z}) = \vec{x} \cdot \vec{y} + \vec{x} \cdot \vec{z} \quad (\text{distributive})$$

$$(k\vec{x}) \cdot \vec{y} = \vec{x} \cdot (k\vec{y}) = k(\vec{x} \cdot \vec{y}) \quad \text{where } k \text{ is a scalar}$$

$$\vec{x} \cdot \vec{y} = 0 \text{ iff } \vec{x} \text{ and } \vec{y} \text{ are orthogonal}$$

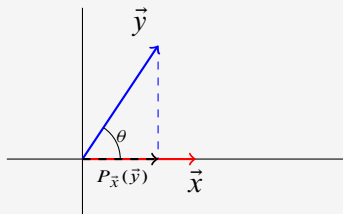
# Vector Projection

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  $P_{\vec{x}}(\vec{y})$ .

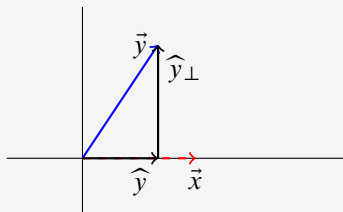
$$C_{\vec{x}}(\vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|} = |\vec{y}| \cos \theta$$



## Vector Projection II

$\vec{y}$  can be decomposed into a vector parallel to  $\vec{x}$ ,  $\hat{y} = P_{\vec{x}}(\vec{y})$ , and 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  $\vec{x}$

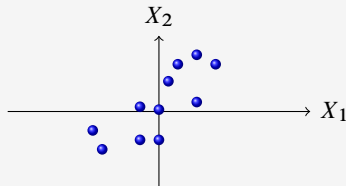
# Vector Geometry of Simple Statistics

# Variable Space Representation of a Data Set

Consider a data set in which we've measured variables  $X_1, X_2, \dots, X_p$ , on a set of subjects (objects)  $a_1, \dots, a_n$ .

	$X_1$	$X_2$
$a_1$	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.





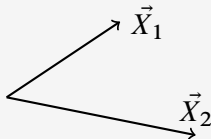
# Subject Space Representation of a Data Set

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

- trickier to visualize because high dimensional

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

- Any pair of non-parallel vectors (of arbitrary dimension) defines a plane.
- Let the variables be represented by centered vectors
  - lengths are proportional to variability
  - angle between vectors represents association or similarity



# Geometry of the Mean in Subject Space

The mean, as you know, is the 'optimal' (in a least square sense) single number summary of a variable of interest.

- The mean,  $\bar{x}$ , minimizes the quantity  $\sum_{i=1}^n (x_i - \bar{x})^2$ .
- The above can be written as  $|\vec{x} - \vec{1}\bar{x}|^2$  where  $\vec{1} = [1, 1, \dots, 1]'$
- We are looking, therefore, for the scalar multiple,  $\bar{x}$ , of the unit vector that minimizes  $|\vec{x} - \vec{1}\bar{x}|^2$

# Geometry of the Mean in Subject Space II

- Recall that:

$$P_{\vec{1}}(\vec{x}) = \vec{1}\bar{x} \text{ for some } \bar{x} \quad (1)$$

$$(\vec{x} - P_{\vec{1}}(\vec{x})) \cdot \vec{1} = 0 \quad (2)$$

- Substituting (1) into (2):

$$(\vec{x} - \vec{1}\bar{x}) \cdot \vec{1} = 0 \quad (3)$$

$$\vec{x} \cdot \vec{1} = \bar{x}(\vec{1} \cdot \vec{1}) \quad (4)$$

- Expanding (4):

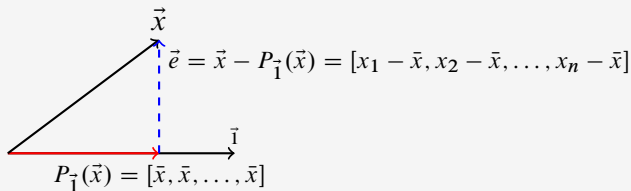
$$x_1 + x_2 + \cdots + x_n = n\bar{x} \quad (5)$$

$$\sum x_i = n\bar{x} \quad (6)$$

$$\bar{x} = (1/n) \sum x_i \quad (7)$$

# Geometry of the Mean in Subject Space III

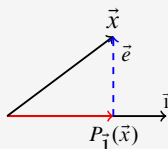
Geometric derivation of the sample mean:



# Geometry of Sample Variance

- $|\vec{e}|^2$  is the sum of squared errors (SSE).
- What is the dimensionality of  $\vec{e}$  ?
  - Because  $\vec{e}$  is orthogonal to the  $n$ -dimensional unit vector  $\vec{1}$ , it must lie in a subspace of dimensionality  $n - 1$ .
- The mean squared error (MSE) is the average error 'per dimension'

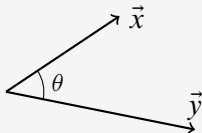
$$\begin{aligned} \text{MSE} &= |\vec{e}|^2 / (n - 1) \\ &= \frac{1}{(n - 1)} \sum (x_i - \bar{x})^2 \leftarrow \text{Sample Variance!} \end{aligned}$$



This is a nice geometric demonstration of why the degrees of freedom of the sample variance is  $n - 1$ .

## Correlation in Vector Geometric Terms

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



$$\text{cor}(X, Y) = r_{XY} = \cos \theta = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|}$$

# Bivariate Regression as Projection

The standard bivariate regression equation relating one observed variable  $X$  (the predictor) to another observed variable of interest,  $Y$  (the outcome) is usually written as:

$$\hat{Y} = a + bX.$$

where  $\hat{Y}$  is the predicted value of  $Y$  and  $a$  and  $b$  are scalar values chosen to minimize  $|Y - \hat{Y}|$ .

Let's express this in vector terms, and work with mean-centered vectors so the equation becomes:

$$\vec{\hat{y}} = b\vec{x}$$

See Wickens Chapt 3 for the general derivation for uncentered variables.

# Bivariate Regression as Projection II

Regression equation for mean-centered vectors:  $\vec{\hat{y}} = b\vec{x}$

- Our goal is to choose the scalar  $b$  such that the error vector  $\vec{e} = \vec{y} - \vec{\hat{y}}$  is as small as possible.
- We've already seen this problem when we derived the mean. We're trying to solve for  $b$  in the equation:

$$\begin{aligned}(\vec{y} - b\vec{x}) \cdot \vec{x} &= 0 \\ \vec{x} \cdot \vec{y} &= b(\vec{x} \cdot \vec{x})\end{aligned}$$

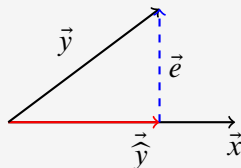
- Solving for  $b$  we get:

$$b = \frac{\vec{x} \cdot \vec{y}}{\vec{x} \cdot \vec{x}} = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2}$$



# Geometry of Bivariate Regression

Geometric interpretation of regression as projection:



We showed on the previous page that

$$\vec{\hat{y}} = b\vec{x} \quad \text{and} \quad b = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|^2}.$$

We can rewrite this in the following ways:

$$\begin{aligned} b &= |x||y| \cos \theta / |x|^2 \\ &= \cos \theta (|y|/|x|) \\ &= r_{XY} (|y|/|x|) \end{aligned}$$

# Bivariate Regression, Goodness of Fit

How well does our prediction agree with our outcome?

- Measure the angle between  $\vec{\hat{y}}$  and  $\vec{y}$ :

$$R = \cos \theta_{\vec{y}, \vec{\hat{y}}} = \frac{|\vec{\hat{y}}|}{|\vec{y}|}$$

- In the single-predictor case  $R = r_{XY}$ , but this is not generally true when we have multiple predictors.
- Note that  $|\vec{y}|$  can be expressed as follows:

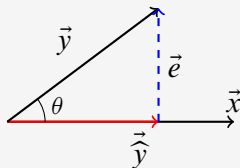
$$\begin{aligned} |\vec{\hat{y}}|^2 + |\vec{e}|^2 &= |\vec{y}|^2 \\ SS_{\text{regression}} + SS_{\text{residual}} &= SS_{\text{total}} \end{aligned}$$

- With simple substitution we can show that:

$$\begin{aligned} SS_{\text{regression}} &= R^2 SS_{\text{total}} \\ SS_{\text{residual}} &= (1 - R^2) SS_{\text{total}} \end{aligned}$$

# Geometry of Goodness of Fit

Geometric interpretation of regression goodness-of-fit:

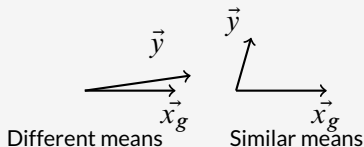


$$R = \cos \theta$$
$$|\vec{\hat{y}}|^2 + |\vec{e}|^2 = |\vec{y}|^2$$

## Two-group ANOVA as Regression

We can also use a geometric perspective to test whether the mean of a variable differs between two groups of subjects.

- Setup a 'dummy variable' as the predictor  $X_g$ . We assign all subjects in group 1 the value 1 and all subjects in group 2 the value -1 on the dummy variable. We then regress the variable of interest,  $Y$ , on  $X_g$ .
- When the means are different in the two groups,  $X_g$  will be a good predictor of the variable of interest, hence  $\vec{y}$  and  $\vec{x}_g$  will have a small angle between them.
- When the means in the two groups are similar, the dummy variable will not be a good predictor. Hence the angle between  $\vec{y}$  and  $\vec{x}_g$  will be large.



# Homework

## ■ Reading

- Wickens, Chapters 1-3
- Kampstra, P. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *J. of Statistical Software*, 28.
- Paradis, E. 2005. R for beginners. Chapter 4.

## ■ Programming exercises

- See handout