

# Scientific Computing for Biologists

## Principal Components Analysis and Eigenanalysis

Instructor: Paul M. Magwene

27 September 2011

# Overview of Lecture

- Principal Components Analysis
  - Variable space representation
  - Subject space representation
  - Mathematical constraints
  - PC scores and loadings
  - Dimension reduction
- Eigenvectors and eigenvalues

# Hands-on Session

- Eigenanalysis and PCA in R
- Eigenanalysis in Python

# General Idea Behind PCA

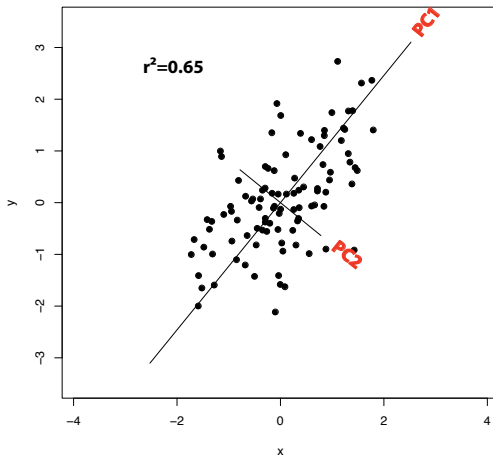
## Goal

Define a new basis for describing your data that has “nice” properties.

By nice we mean:

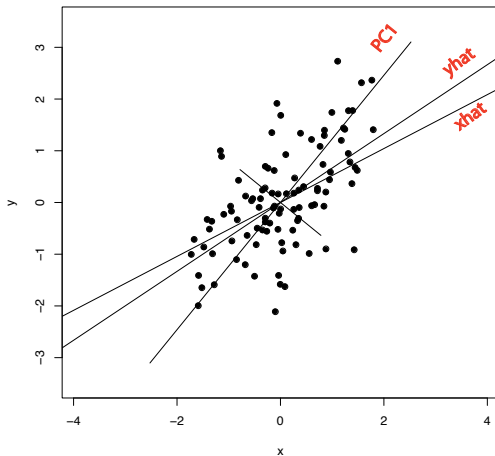
- spans same space as original basis
- provides an orthogonal basis
- the order of the basis vectors is related to their relative importance
- can be used to facilitate low dimensional summaries of high dimensional data

## Example PC Basis

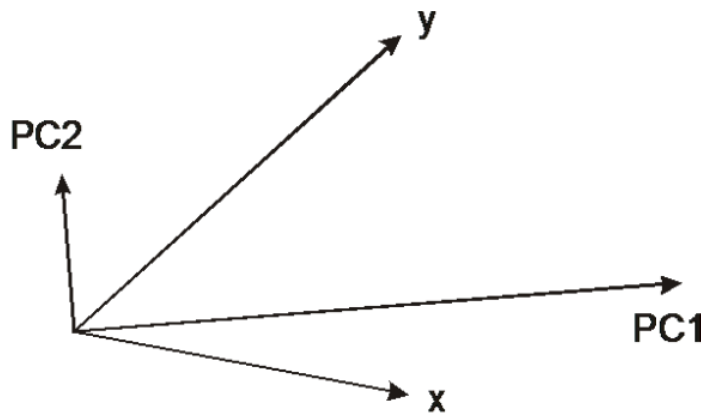


Variable Space Representation

# Contrast Between PC Axes and Regression



## Subject Space Representation of PCA



Again  $r^2 = 0.65$ .

# Mathematical Constraints on PCA

The principal components are linear combinations of the original variables

$$\begin{aligned}u_1 &= a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\u_2 &= a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\&\vdots \\u_p &= a_{1p}x_1 + a_{2p}x_2 + \cdots + a_{pp}x_p\end{aligned}$$

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_p \end{bmatrix}$$

In this formulation each of the  $u_i$  represent column vectors.



## Mathematical Constraints on PCA, cont.

- The sum of the squared coefficients for each PC is fixed to unity:

$$a_{1k}^2 + a_{2k}^2 + \cdots + a_{pk}^2 = 1$$

- The sum of the squared lengths for the original variables and the PCs is the same:

$$|x_1|^2 + |x_2|^2 + \cdots + |x_p|^2 = |u_1|^2 + |u_2|^2 + \cdots + |u_p|^2$$

- The PCs are orthogonal:

$$a_j \cdot a_k = 0 \text{ for all } j \neq k.$$

- The PCs are ordered such that:

$$|u_1|^2 \geq |u_2|^2 \geq \cdots \geq |u_p|^2$$

# Principal Component Scores

## Definition

The PC scores are the components of the original observations with respect to the principal components (i.e. the observations represented in the new basis).

If the values of the  $i$ -th individual in the original variables are:

$$\mathbf{r}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$$

then the score of that individual on the  $k$ -th principal component axis is given by:

$$\mathbf{r}_i \cdot \mathbf{a}_k = a_{1k}x_{i1} + a_{2k}x_{i2} + \dots + a_{pk}x_{ip}$$

# Principal Component Loadings

The **loading vector** of the  $k$ -th PC is:

$$|u_k|a_k$$

The **loading** of the  $j$ -th original variable on the  $k$ -th PC is:

$$|u_k|a_{jk}$$

## Interpretation

Loadings give a sense of the relative contribution of each of the original variables to the PC of interest.

# PCA Cautions and Caveats

## Scale matters

If the original variables are not measured on comparable scales they should be standardized prior to PCA

## PCA emphasizes correlated variables

If one of the original variables is (nearly) independent of the other variables than it will have weak loadings on the major PCs. But this doesn't mean it's unimportant!

## PCA does not represent a model

It's simply a transformation of the original data.

# PCA for Dimension Reduction

A common application of PCA is to find a lower dimensional representation of a high dimensional data set.

## Goal

Capture the majority of the variance with just a few principal component axes.

# PCA Example: Jolicoeur and Mosimann's Turtle Data Set

- 48 specimens ( $\sigma, \varphi$ ), 3 variables (length, width, height)
- Covariance and correlation matrices:

$$\text{cov}(X) = \begin{bmatrix} 420 & 254 & 165 \\ & 161 & 102 \\ & & 70 \end{bmatrix}; \text{cor}(X) = \begin{bmatrix} 1 & 0.978 & 0.964 \\ & 1 & 0.961 \\ & & 1 \end{bmatrix}$$

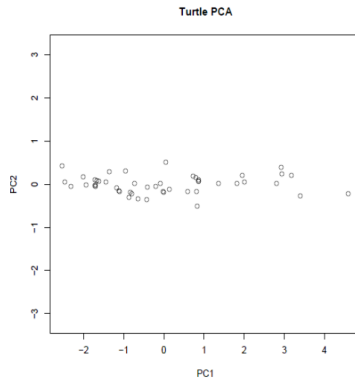
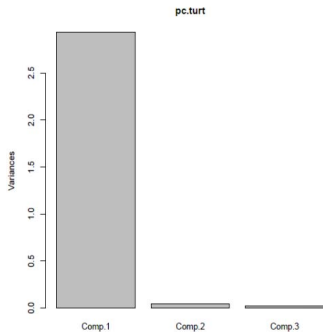
- Principal components (using correlation matrix), proportion of variance:

PC1	PC2	PC3
0.978	0.014	0.007

- Principal component loadings:

	PC1	PC2	PC3
L	0.579	-0.325	0.748
W	0.578	-0.483	-0.657
H	0.575	0.813	~0

# Turtle PCA Plots



# Eigenvectors and Eigenvalues



# Reminder about Linear Transformations

Recall:

- Every linear transformation can be represented by a matrix ...
- ... and conversely, every matrix represents a linear transformation
- An  $n \times p$  matrix represents a transformation from  $\mathbb{R}^p$  to  $\mathbb{R}^n$ .
- A  $p \times p$  matrix is special in that it represent a transformation from  $\mathbb{R}^p$  to  $\mathbb{R}^p$

# Linear Transformations and Eigenvectors/Eigenvalues

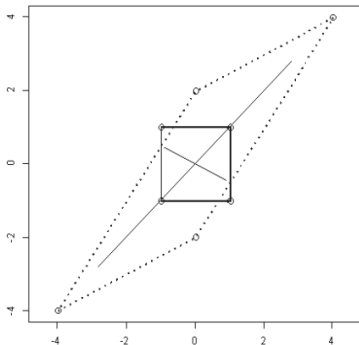
If  $A$  is a  $p \times p$  matrix, then:

- The **eigenvectors** of  $A$  represent directions in  $\mathbb{R}^p$  that are unchanged by the transformation represented by  $A$
- Vectors along the directions defined by the eigenvectors may be scaled however. The relative scaling is given by the **eigenvalues** of  $A$

if  $Av = kv$  for some scalar  $k$  than:

- $v$  is an eigenvector of  $A$
- $k$  is it's associated eigenvalue.

# Example Linear Transformation



$$A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{bmatrix} 2x+2y \\ x+3y \end{bmatrix}$$

$$V = \begin{bmatrix} -0.71 & -0.89 \\ -0.71 & -0.45 \end{bmatrix} \begin{array}{l} \text{Eigenvectors} \\ \text{in} \\ \text{columns} \end{array}$$

$$Q = \begin{bmatrix} 4 & 1 \end{bmatrix} \begin{array}{l} \text{Corresponding} \\ \text{Eigenvalues} \end{array}$$

Note that the eigenvectors above are *not* orthogonal.

## More on Eigenvectors and Eigenvalues

- Eigenvectors/eigenvalues are only defined for square matrices
- Eigenvectors are not, in general, orthogonal

However, if  $A$  is a real symmetric matrix then:

- eigenvectors of  $A$  are guaranteed to be orthogonal
- eigenvalues of  $A$  are guaranteed to be real
- $A$  has zero-valued eigenvectors *iff*  $A$  is singular
- if  $V$  is the matrix of eigenvectors we can write:

$$V^{-1}AV = D$$

where  $D$  is a diagonal matrix with eigenvalues on the diagonal:

$$D = \begin{bmatrix} l_1 & & 0 \\ & l_2 & \\ 0 & & l_3 \end{bmatrix}$$

# Calculation of PCs

Principal components can be calculated by eigenanalysis of the covariance or correlation matrix.

$$\text{let } X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \begin{array}{l} \text{data matrix w/ variables in} \\ \text{columns} \end{array}$$

$$S = \text{cov}(X)$$

$(p \times p)$

$$V = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \dots & \vec{v}_p \end{bmatrix} \quad \begin{array}{l} \text{matrix w/ eigenvectors of } S \\ \text{columns} \end{array}$$

$$\vec{l} = [l_1 \quad l_2 \quad \dots \quad l_p] \quad \text{eigenvalues of } S$$

Variance summarized by  $k^{\text{th}}$  PC:  $\sqrt{l_k}$

PC scores matrix:  $W = X V$

PC loadings matrix:  $V L^{-1/2}$

$$\text{where } L^{-1/2} = \begin{bmatrix} \sqrt{l_1} & & 0 \\ & \sqrt{l_2} & \\ 0 & & \ddots \\ & & & \sqrt{l_p} \end{bmatrix}$$