Scientific Computing for Biologists

Principal Components Analysis and Eigenanalysis

Instructor: Paul M. Magwene

30 September 2014

Overview of Lecture

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

Hands-on Session

- Eigenanalysis and PCA in Python
- Introduction to Biplots

Reminder: Definition of Basis

Let S be a subspace of $\mathbb{R}^n$. Then there is a finite list, $\boldsymbol{X}$, of vectors from S such that S is the space spanned by X.

Let S be a subspace of $\mathbb{R}^n$ spanned by the list $(\vec{\boldsymbol{u}}_1, \vec{\boldsymbol{u}}_2, \ldots, \vec{\boldsymbol{u}}_n)$. Then there is a linearly independent sublist of $(\vec{\boldsymbol{u}}_1, \vec{\boldsymbol{u}}_2, \ldots, \vec{\boldsymbol{u}}_n)$ that also spans S.

Definition

A list X is a **basis** for S if:

- X is linearly independent
- S is the subspace spanned by X

General Idea Behind PCA

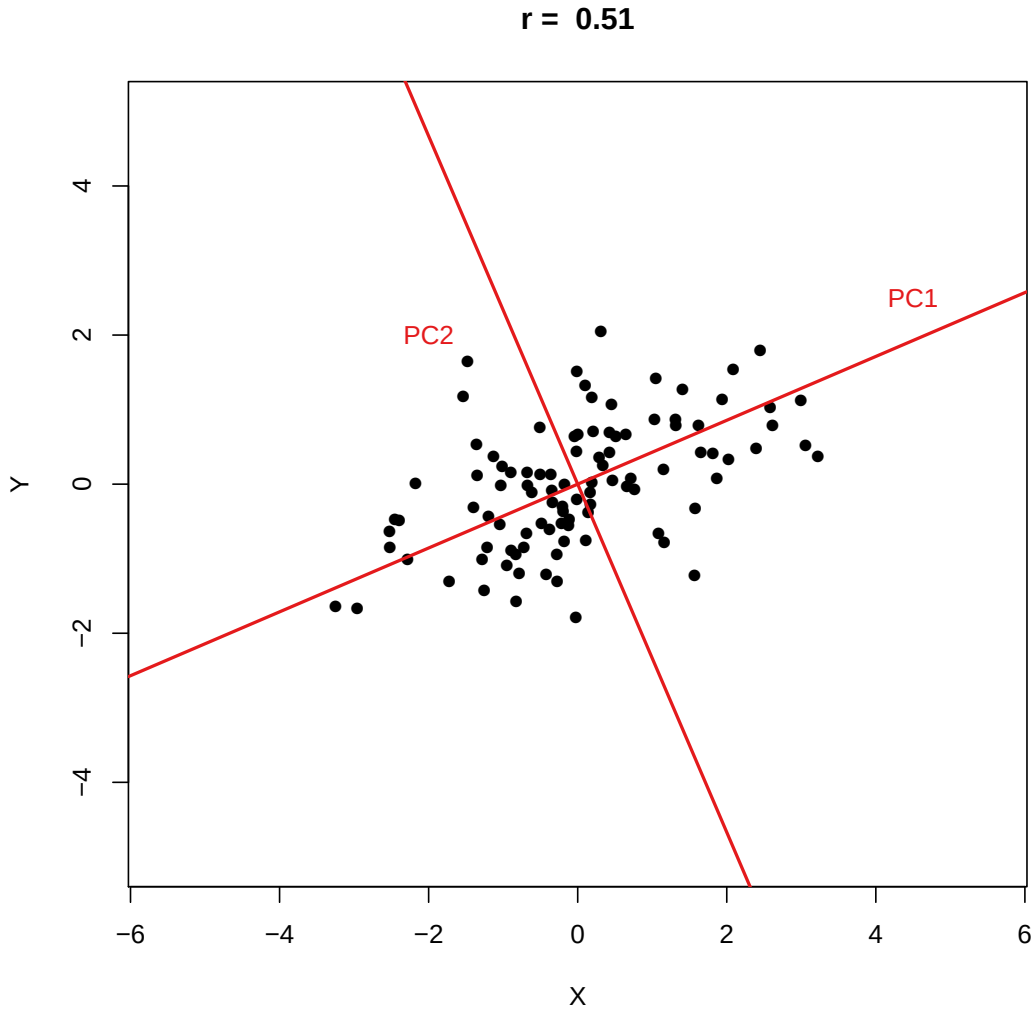
Goal

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

By nice we mean:

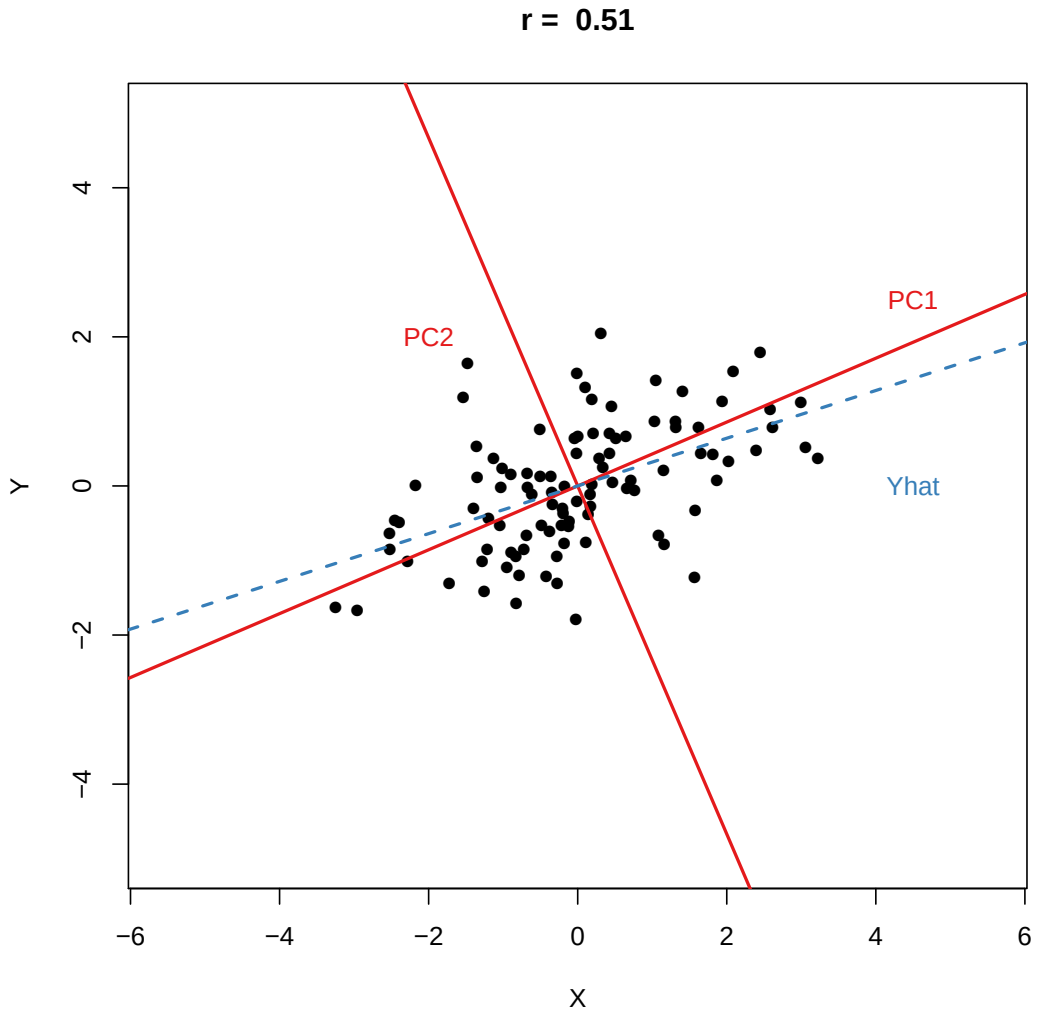
- spans same space as original basis
- provides an orthogonal basis
- the order of the basis vectors is related to their relative “importance"
- can be used to facilitate low dimensional summaries of high dimensional data

Example PC Basis

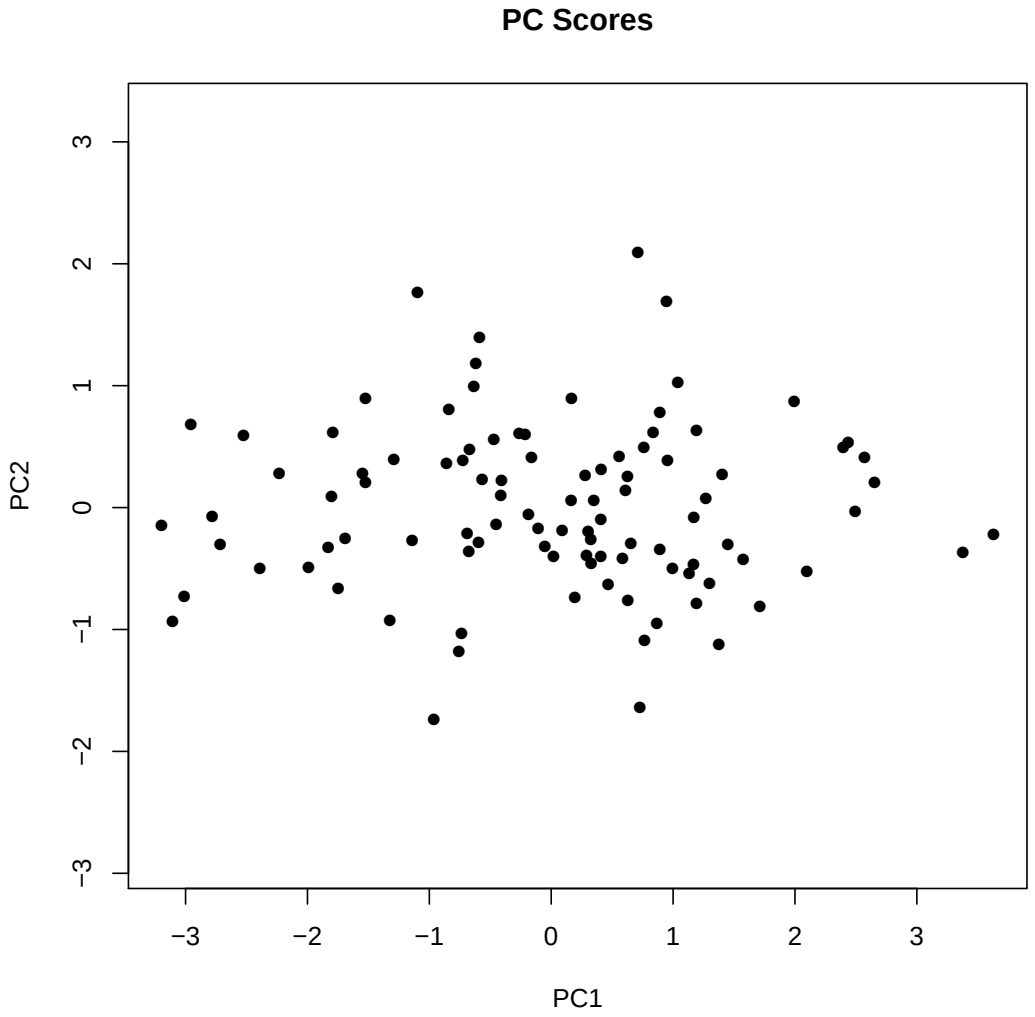


Variable Space Representation

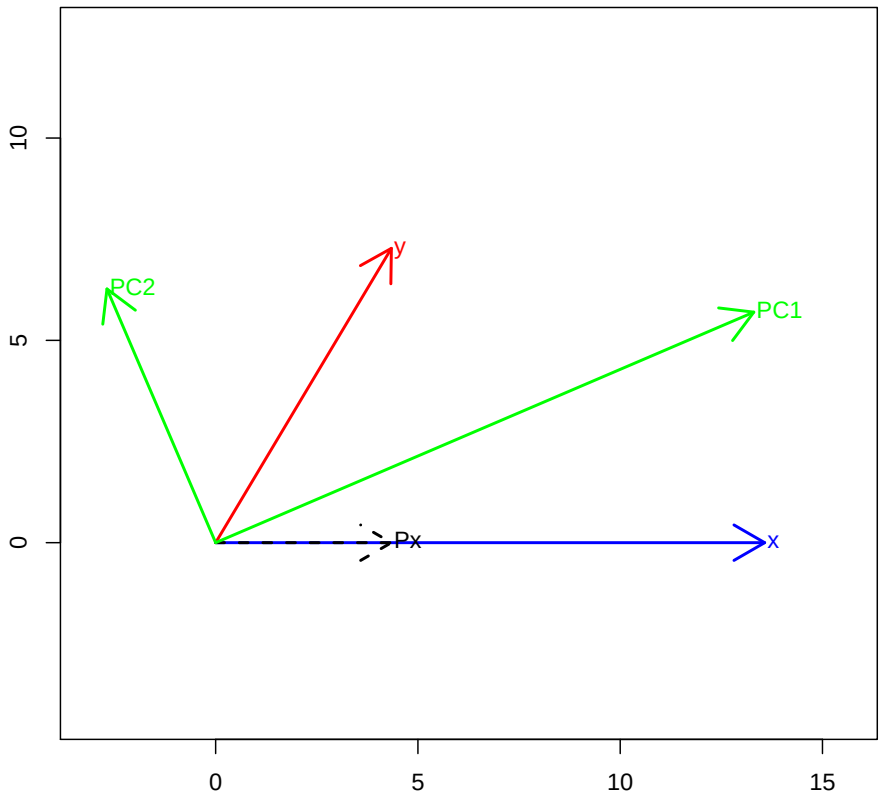
Contrast Between PC Axes and Regression



Observations with Respect to the PC Basis



Subject Space Representation of PCA



Mathematical Constraints on PCA

The principal components are linear combinations of the original variables

$$\begin{aligned}
\vec{u}_1 &= a_{11}\vec{x}_1 + a_{21}\vec{x}_2 + \cdots + a_{p1}\vec{x}_p \\
\vec{u}_2 &= a_{12}\vec{x}_1 + a_{22}\vec{x}_2 + \cdots + a_{p2}\vec{x}_p \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
\vec{u}_p &= a_{1p}\vec{x}_1 + a_{2p}\vec{x}_2 + \cdots + a_{pp}\vec{x}_p
\end{aligned}$$

In this formulation the $\vec{x}_i$ and $\vec{u}_i$ are column vectors.

Mathematical Constraints on PCA, cont.

$$\underset{(n\times p)}{\boldsymbol{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{p2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \vec{\boldsymbol{x}}_1 & \vec{\boldsymbol{x}}_2 & \cdots & \vec{\boldsymbol{x}}_p \end{bmatrix}$$

$$\underset{(p\times p)}{\boldsymbol{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} = \begin{bmatrix} \vec{\boldsymbol{a}}_1 & \vec{\boldsymbol{a}}_2 & \cdots & \vec{\boldsymbol{a}}_p \end{bmatrix}$$

$$\underset{(n\times p)}{\boldsymbol{U}} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{n1} & u_{p2} & \cdots & u_{np} \end{bmatrix} = \begin{bmatrix} \vec{\boldsymbol{u}}_1 & \vec{\boldsymbol{u}}_2 & \cdots & \vec{\boldsymbol{u}}_p \end{bmatrix} = \boldsymbol{XA}$$

Mathematical Constraints on PCA, cont.

- The PCs are orthogonal:

$$\vec{a}_j \cdot \vec{a}_k = 0 \text{ for all } j \neq k.$$

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

$$\vec{a}_k \cdot \vec{a}_k = 1$$

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

$$|\vec{x}_1|^2 + |\vec{x}_2|^2 + \cdots + |\vec{x}_p|^2 = |\vec{u}_1|^2 + |\vec{u}_2|^2 + \cdots + |\vec{u}_p|^2$$

- The PCs are ordered such that:

$$|\vec{u}_1|^2 \geq |\vec{u}_2|^2 \geq \cdots \geq |\vec{u}_p|^2$$

Principal Component Scores

Definition

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

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

$$\boldsymbol{r}_i \quad = \quad [x_{i1}, \quad x_{i2}, \quad \cdots \quad x_{ip}]$$

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

$$\boldsymbol{U}_{ik} = \boldsymbol{r}_i \cdot \boldsymbol{a}_k = a_{1k}x_{i1} + a_{2k}x_{i2} + \cdots + a_{pk}x_{ip}$$

Principal Component Loadings

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

$$|\boldsymbol{u}_k|\boldsymbol{a}_k$$

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

$$|\boldsymbol{u}_k|a_{jk}$$

Interpretation

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

Bivariate Example Revisited

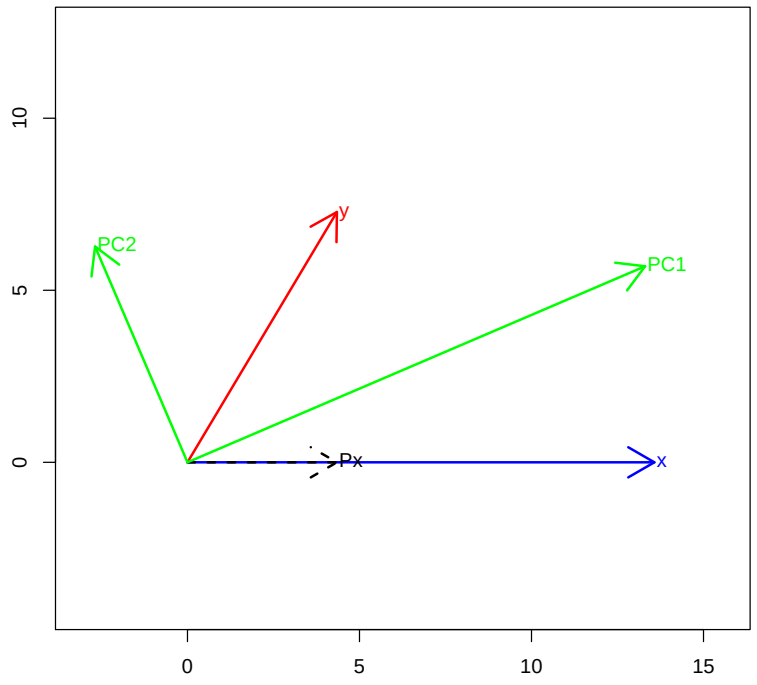
```
> cov(z)
         [,1]      [,2]
[1,] 1.865551 0.6523660
[2,] 0.652366 0.9248509

> z.pca
Standard deviations:
[1] 1.4830530 0.7687363

Rotation:
             PC1         PC2
[1,] -0.8901780 -0.4556128
[2,] -0.4556128  0.8901780

> A <- z.pca$rotation # coefficients
> L <- diag(z.pca$sdev) # lengths of PCs
> A %*% L  # loadings
          [,1]       [,2]
[1,] -1.320181 -0.3502461
[2,] -0.675698  0.6843122
```

```
> (A %*% L)**2  # loadings squared
          [,1]      [,2]
[1,] 1.7428784 0.1226723
[2,] 0.4565677 0.4682832
> apply(z, 2, var) # variance of orig.
[1] 1.8655507 0.9248509
```



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 (♂,♀), 3 variables (length, width, height)
- Covariance and correlation matrices:

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

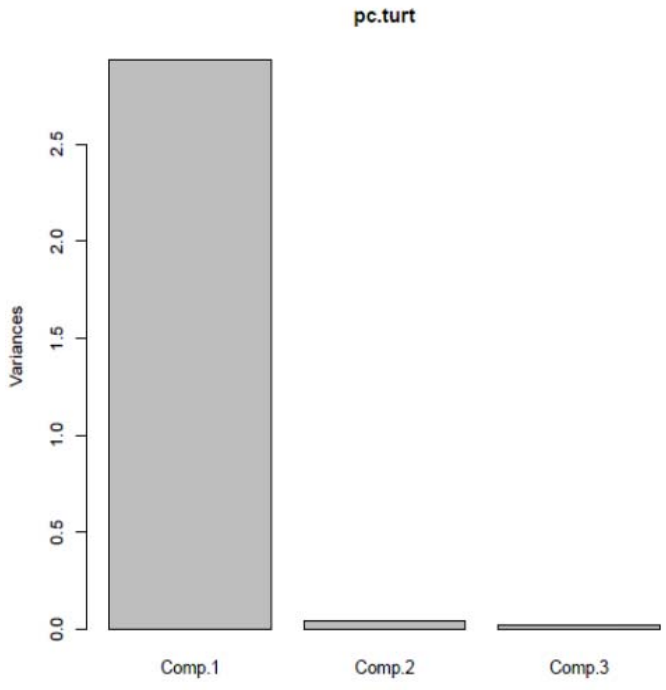
- Principal components (using correlation matrix), proportion of variance:

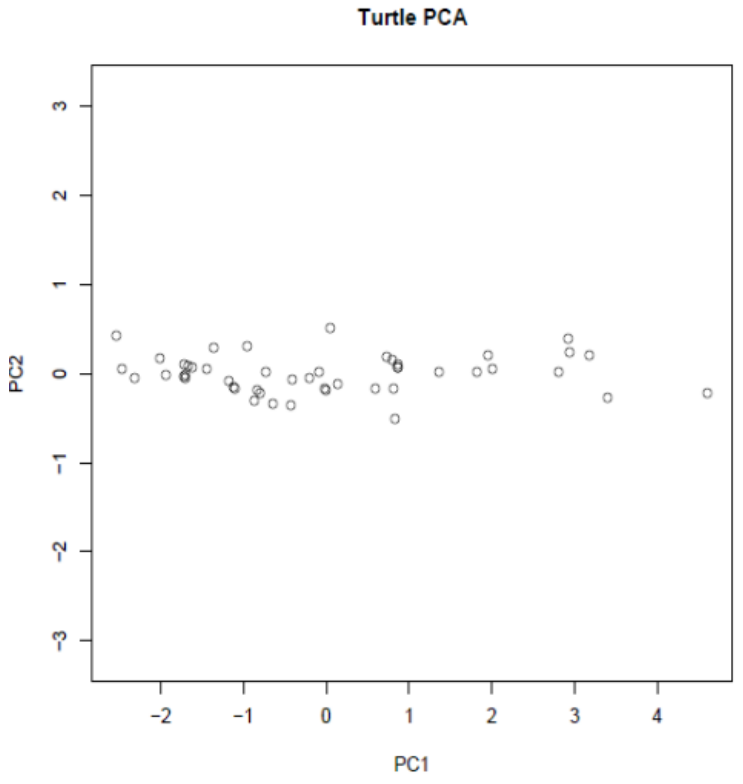
PC1	PC2	PC3
0.978	0.014	0.007

- 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	0.092

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

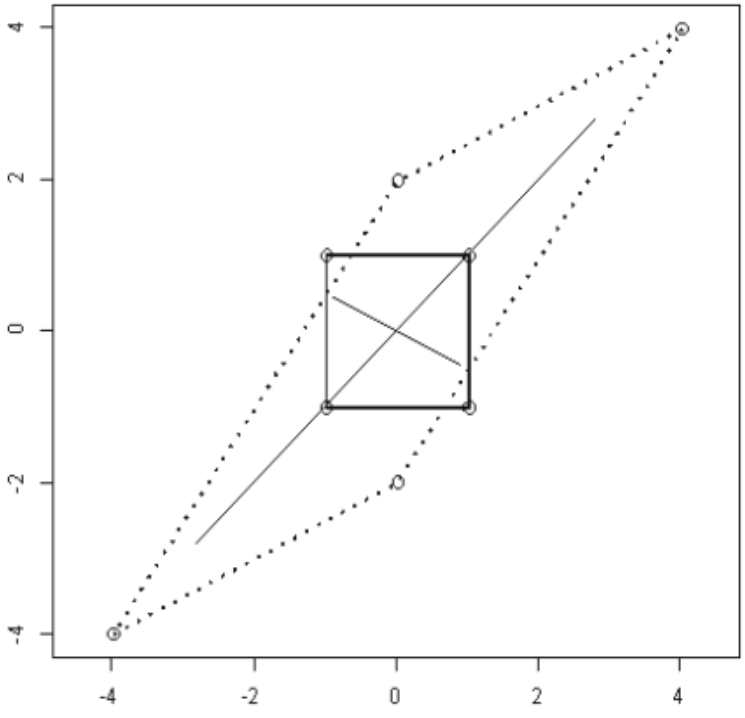
If $\boldsymbol{A}$ is a $p \times p$ matrix, then:

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

if $\boldsymbol{A}\boldsymbol{v} = k\boldsymbol{v}$ for some scalar k than:

- $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{A}$
- k is it's associated eigenvalue.

Example Linear Transformation



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

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

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

$$\ell = [\,4 \qquad 1\,]$$ Corresponding Eigenvalues

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 $\boldsymbol{A}$ is a real symmetric matrix then:

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

$$\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D}$$

where $\boldsymbol{D}$ is a diagonal matrix with eigenvalues on the diagonal:

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

Calculation of PCs

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

let $\underset{(n \times p)}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & & x_{np} \end{bmatrix}$ data matrix w/ variables in columns

$\underset{(p \times p)}{S} = \text{cov}(X)$

$\underset{(p \times p)}{V} = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_p \end{bmatrix}$ matrix w/ eigenvectors of S columns

$\vec{l} = \begin{bmatrix} l_1 & l_2 & \cdots & l_p \end{bmatrix}$ eigenvalues of S

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

PC scores matrix: $W = XV$

PC loadings matrix: $VL^{-\frac{1}{2}}$ where $L^{-\frac{1}{2}} = \begin{bmatrix} \sqrt{l_1} & & 0 \\ & \sqrt{l_2} & \\ 0 & & \sqrt{l_p} \end{bmatrix}$