Scientific Computing for Biologists
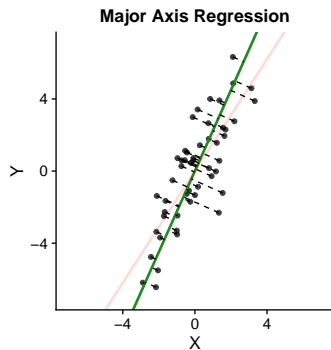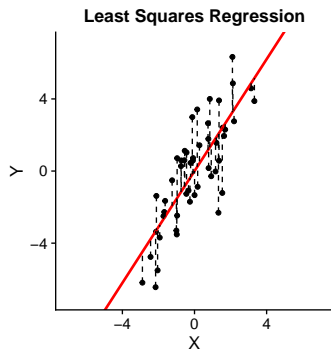
## Principal Components Analysis and Eigenanalysis

Paul M. Magwene

Overview of Lecture

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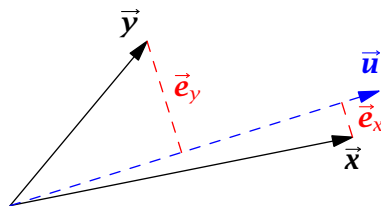- Eigenvectors and eigenvalues

Least Squares Regression vs. Major Axis Regression

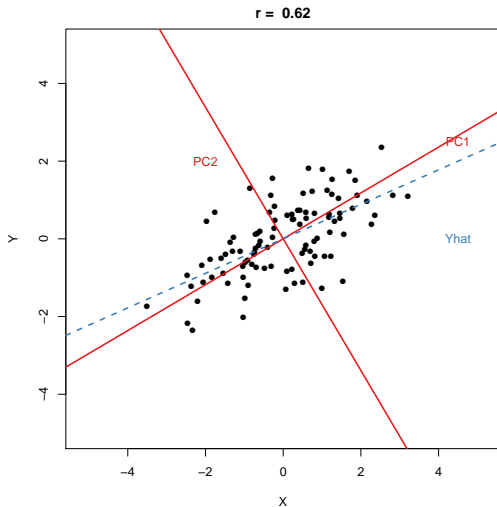# Vector Geometry of Least Squares vs. Major Axis Regression
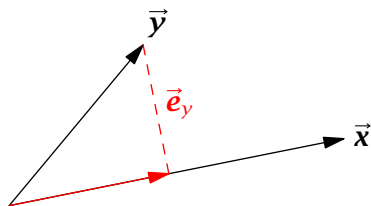
(a) OLS   (b) Major Axis Regression

Figure: Vector geometry of ordinary least-squares and major axis regression.

Contrast Between PC Axes and Regression

Observations with Respect to the PC Basis
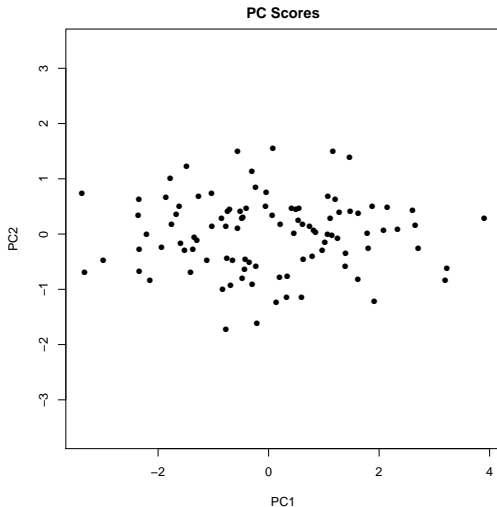
Subject Space Representation of PCA
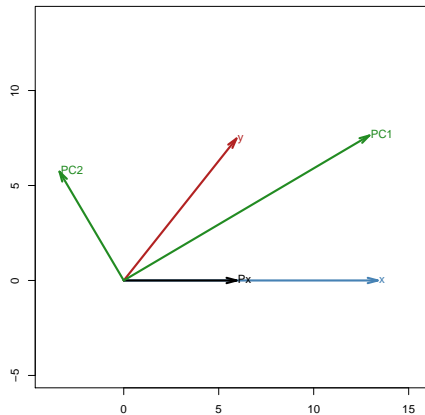
General Idea Behind PCA

> ## 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

Space Spanned by a List of Vectors

> ## Definition
>
> Let X be a finite list of n-vectors. The **space spanned** by X is the set of all vectors that can be written as linear combinations of the vectors in X.
>
> A space spanned includes the zero vector and is closed under addition and multiplication by a scalar.

Remember that a *linear combination* of vectors is an equation of the form $z = b_1 \boldsymbol{x}_1 + b_2 \boldsymbol{x}_2 + \cdots + b_p \boldsymbol{x}_p$

Subspaces

$\mathbb{R}^n$ denotes the seat of real *n*-vectors - the set of all $n \times 1$ matrices with entries from the set $\mathbb{R}$ of real numbers.

Definition

A **subspace** of $\mathbb{R}^n$ is a subset S of $\mathbb{R}^n$ with the following properties:

1. $0 \in S$
2. If $\boldsymbol{u} \in S$ then $k\boldsymbol{u} \in S$ for all real numbers k
3. If $\boldsymbol{u} \in S$ and $\boldsymbol{v} \in S$ then $\boldsymbol{u} + \boldsymbol{v} \in S$

Examples of subspaces of $\mathbb{R}^n$:

- any space spanned by a list of vectors in $\mathbb{R}^n$
- the set of all solutions to an equation $A\boldsymbol{x} = 0$ where A is a $p \times n$ matrix, for any number p.

Basis

Let S be a subspace of $\mathbb{R}^n$. Then there is a finite list, 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 $(u_1, u_2, \ldots, u_n)$. Then there is a linearly independent sublist of $(u_1, u_2, \ldots, u_n)$ 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

Dimension

Let S be a subspace of $\mathbb{R}^n$.

> **Definition**
>
> The **dimension** of S is the number of elements in a basis for S.

Mathematical Constraints on PCA

The principal components are linear combinations of the original variables

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

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

Mathematical Constraints on PCA, cont.

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

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

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

Mathematical Constraints on PCA, cont.

- The PCs are orthogonal:

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

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

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

- 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:

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

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

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

Principal Component Loadings

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

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

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

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

Interpretation

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

- When the input data is unscaled, loadings are the covariances between the original variables and the PCs
- When the input data are scaled (i.e. sd = 1), loading are the correlations between the original variables and the PCs

PCA Example: Jolicoeur and Mosimann's Turtle Data Set

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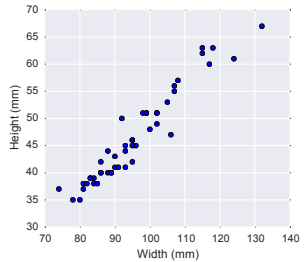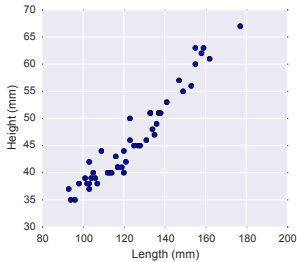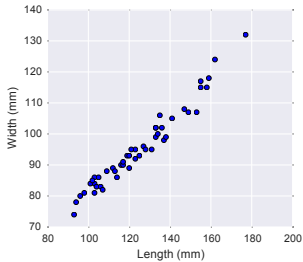
PCA Example: Jolicoeur and Mosimann's Turtle Data Set

- 48 specimens (σ, φ), 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	0.092

Turtle PCA Plots
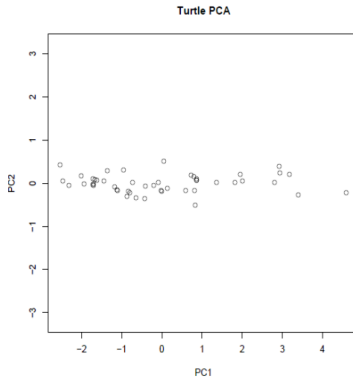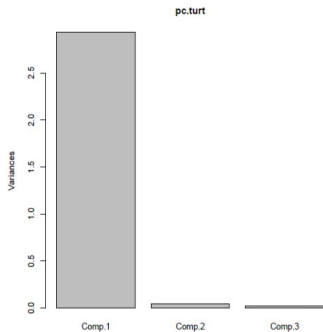
PCA Cautions and Caveats

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.

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.

Eigenvectors and Eigenvalues

Matrices as Linear Transformations

- Let A be a particular $n \times p$ matrix. Than for any p-vector $\boldsymbol{x}$, the product $A\boldsymbol{x}$ is a n-vector.
- We say that the matrix A determines a function from $\mathbb{R}^p$ to $\mathbb{R}^n$.
    - $A(k\boldsymbol{x}) = k(A\boldsymbol{x})$ where k is a scalar.
    - If $\boldsymbol{y}$ is also a p-vector than $A(\boldsymbol{x} + \boldsymbol{y}) = A\boldsymbol{x} + A\boldsymbol{y}$ is an n-vector
- A function, f, where $f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$ and $f(k\boldsymbol{x}) = kf(\boldsymbol{x})$ is called a ***linear transformation***.

Highlight

Every matrix determines a linear transformation!

Every linear transformation can be represented by a matrix!

Examples of Linear Transformation in $\mathbb{R}^2$

- reflection in the x-axis

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x \\ -y \end{bmatrix}$$

- reflection in the line $y = x$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} y \\ x \end{bmatrix}$$

- shear parallel to the x-axis

$$\begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x + ay \\ y \end{bmatrix}$$
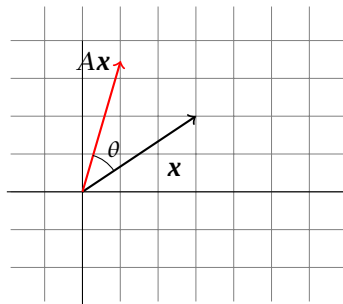
- projection onto the x-axis

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x \\ 0 \end{bmatrix}$$

- How about reflection in the y-axis? shear parallel to the y-axis? projection onto the y-axis?

Examples of Linear Transformation: Rotation

- The rotation of the plane, by an angle θ about the origin is given by:

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

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

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

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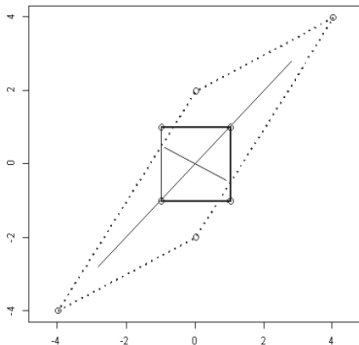
$$A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}$$

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

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

$$\ell = \begin{bmatrix} 4 & 1 \end{bmatrix} \text{ Corresponding Eigenvalues}$$

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 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.

$$\underset{(n \times p)}{\boldsymbol{X}} = \begin{bmatrix} \vec{\boldsymbol{X}}_1 & \vec{\boldsymbol{X}}_2 & \cdots & \vec{\boldsymbol{X}}_p \end{bmatrix}$$ Data matrix w/variables in columns

$$\underset{(p \times p)}{\boldsymbol{S}} = \text{cov}(X)$$ Covariance or correlation matrix of X

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

$$\underset{(1 \times p)}{\boldsymbol{l}} = \begin{bmatrix} l_1 & l_2 & \cdots & l_p \end{bmatrix}$$ Vector of eigenvalues of S

Given the above:

- PC scores matrix: $\underset{(p \times p)}{\boldsymbol{U}} = \boldsymbol{X}\boldsymbol{V}$

- PC loadings matrix: $\boldsymbol{V}\boldsymbol{L}^{1/2}$ where $\boldsymbol{L}^{1/2} = \begin{bmatrix} \sqrt{l_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{l_p} \end{bmatrix}$

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