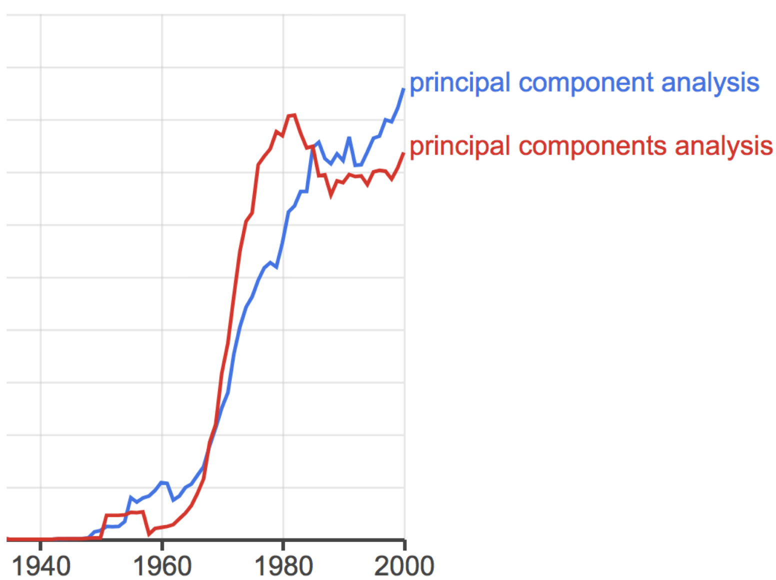
# Principal Component Analysis (PCA): Part I

CLPS 2908
Lecture 10
Mar 5, 2019

# The Name

Google *ngram*



# Linear Combinations

| | X1 | X2 | X3 | X4 | M | SUM | LC1 | LC2 | LC3 | PC1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 3 | 5 | 4 | 4 | 16 | 3.1 | 3.9 | -0.5 | 0.77 |
| | 8 | 6 | 2 | 2 | 4.5 | 18 | 3.5 | 6.0 | 2.5 | -2.81 |
| | 6 | 5 | 7 | 5 | 5.75 | 23 | 4.5 | 5.8 | -0.25 | 0.53 |
| | 4 | 5 | 5 | 7 | 5.25 | 21 | 3.8 | 4.8 | -0.75 | 1.35 |
| M | 5.5 | 4.75 | 4.75 | 4.5 | 4.9 | 19.5 | 3.7 | 5.1 | 0.3 | 0.0 |
| $s^2$ | **3.67** | **1.58** | **4.25** | **4.33** | **0.60** | **9.67** | **0.34** | **0.95** | **2.29** | **3.54** |

**LC1 = (X1*1 + X2*0.5 + X3*1 + X4*0.5)/4**

**LC2 = (X1*2 + X2*1 + X3*0.5 + X4*0.5)/4**

**LC3 = (X1*1 + X2 *1 + X3 *(-1) + X4 *(-1))/4**

**PC1 = X1*(-0.329)+X2*(-0.241)+X3*(0.276)+X4*(0.356)**

# Linear Combinations

| | X1 | X2 | X3 | X4 | M | SUM | LC1 | LC2 | LC3 | PC1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 3 | 5 | 4 | 4 | 16 | 3.1 | 3.9 | -0.5 | 0.77 |
| | 8 | 6 | 2 | 2 | 4.5 | 18 | 3.5 | 6.0 | 2.5 | -2.81 |
| | 6 | 5 | 7 | 5 | 5.75 | 23 | 4.5 | 5.8 | -0.25 | 0.53 |
| | 4 | 5 | 5 | 7 | 5.25 | 21 | 3.8 | 4.8 | -0.75 | 1.35 |
| M | 5.5 | 4.75 | 4.75 | 4.5 | 4.9 | 19.5 | 3.7 | 5.1 | 0.3 | 0.0 |
| $s^2$ | **3.67** | **1.58** | **4.25** | **4.33** | **0.60** | **9.67** | **0.34** | **0.95** | **2.29** | **3.54** |

**LC1 = (X1*1 + X2*0.5 + X3*1 + X4*0.5)/4**

**LC2 = (X1*2 + X2*1 + X3*0.5 + X4*0.5**

**LC3 =X1*1 + X2 *1 + X3 *(-1) + X4 *(-1)**

**PC1 = X1*(-0.329)+X2*(-0.241)+X3*(0.276)+X4*(0.356)**

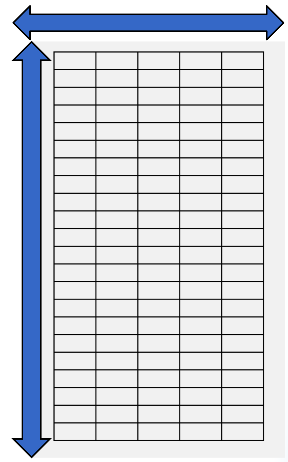
# PCA's Major Elements

**Focus I**:  Subject variance = distinguishing information

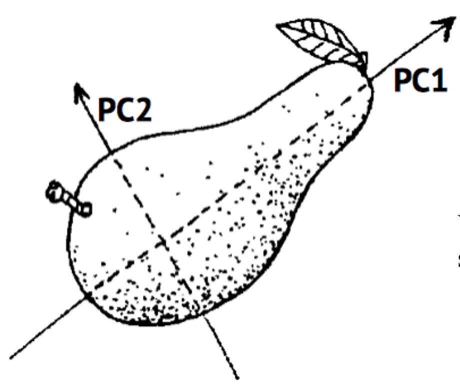
**Focus II:** Variables' covariance

**Goal of PCA**:  Capture much of the variance with fewer variates (components) than the number of original variables.

**Constraints**:  Let components be orthogonal; let each have (relative) maximal variance.



# Pear Component Analysis

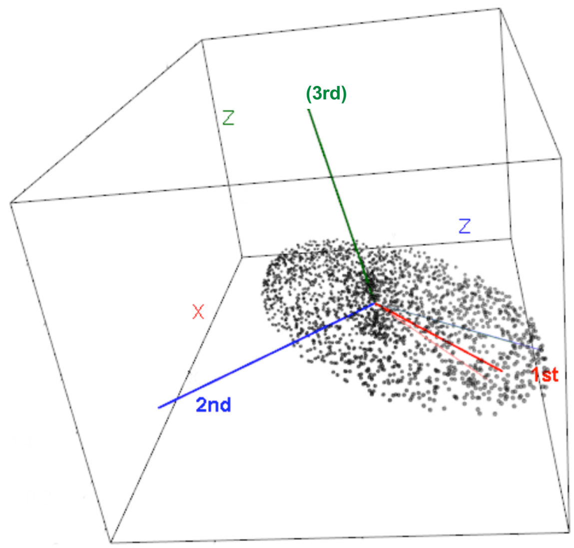



Volume of pear = swarm of variables

Figure 7.94: Pear on the sword: PCA metaphor.

Shipunov, Alexey (and many others). Visual statistics. Use R!
May 4, 2017 version. 388 pp.  URL: http://ashipunov.info/shipunov/school/biol_240/en/

# A clear case




# Variances of Linear Combinations

Linear combination: $\mathbf{C} = \mathbf{X}\mathbf{w}$

$$\mathbf{S_{cc}} = \mathbf{C'C} \cdot (n-1)^{-1}$$

$$= (\mathbf{Xw})'\mathbf{Xw} \cdot (n-1)^{-1}$$ ← Substitute **X**w for **C**

$$= \mathbf{w}'\underbrace{\boxed{\mathbf{X'X}}\mathbf{w} \cdot \boxed{(n-1)^{-1}}}_{\mathbf{S_{xx}}}$$

PCA can be performed either on the original variance-covariance matrix $\mathbf{S}_{xx}$ or the correlation matrix $\mathbf{R}_{xx}$

$$\mathbf{S_{cc}} = \mathbf{w'S_{xx}w}$$

similarly: $\mathbf{R_{cc}} = \mathbf{v'R_{xx}v} = \mathbf{C'C} \cdot (n-1)^{-1}$,

whereby $\mathbf{C} = \mathbf{Zv}$

# Finding the Components

$$\underset{p\times n}{\mathbf{Z'}} \cdot \underset{n\times p}{\mathbf{Z}} \rightarrow \underset{p\times p}{\mathbf{R}} \begin{bmatrix} s_1^2 & & & \\ r_{21} & s_2^2 & & \\ r_{31} & r_{32} & s_3^2 & \\ r_{41} & r_{42} & r_{43} & s_4^2 \end{bmatrix}$$

"Spectral decomposition": Finding eigenstructures of **R**

**eigenvalues**

$\mathbf{V'}\ \mathbf{R}\ \mathbf{V} = \Lambda$

$r$ = correlation among variables
But also number of components

$\Lambda$ is a var-cov matrix

**eigenvectors** $v_i$ are <u>weights</u> that transform the $p$ variables $X_i$ into $r$ linear combinations
**eigenvalues** $\lambda_j$ are the <u>variances</u> of these $r$ linear combinations.

---

$$\underset{n\times p}{\mathbf{X}} \cdot \underset{p\times r}{\mathbf{V}} = \underset{n\times r}{\mathbf{C}} \qquad \underset{r\times n}{\mathbf{C'}} \cdot \underset{n\times r}{\mathbf{C}} = \underset{r\times r}{\Lambda}$$

**Fundamental equation for PCA:**

$$\underset{r\times p}{\mathbf{V'}}\ \underset{p\times n}{\mathbf{X'}}\ \underset{n\times p}{\mathbf{X}}\ \underset{p\times r}{\mathbf{V}} = \mathbf{V'}\ \underset{p\times p}{\mathbf{R}}\ \mathbf{V} = \underset{r\times r}{\Lambda}$$

**Constraint:** $v'v = 1$

# Origin of Procedure

**Theorem:** For any symmetric $p \times p$ matrix (e.g., **R'** = **R**), there are $p$ "eigenvectors" that are "invariant under transformation by their matrix":

$\mathbf{R}\ v = \lambda_j\ v$ **or** $\mathbf{R}\ \mathbf{V} = \Lambda\ \mathbf{V} \Rightarrow \mathbf{R} = \mathbf{V}\ \Lambda\ \mathbf{V'}$

( = spectral decomposition of R
into eigenstructures)

$\mathbf{R}\ v - \lambda_j\ v = \mathbf{0}$

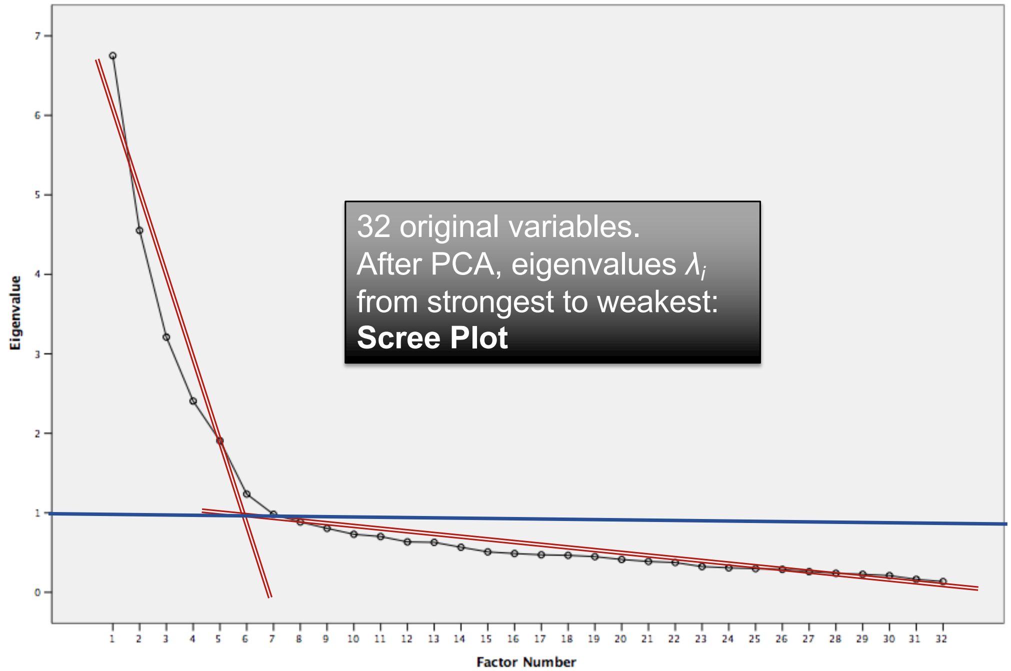
$(\mathbf{R} - \mathbf{I}\ \lambda_j)\ v = \mathbf{0}$

To maximize $\lambda_j$, set the determinant to zero

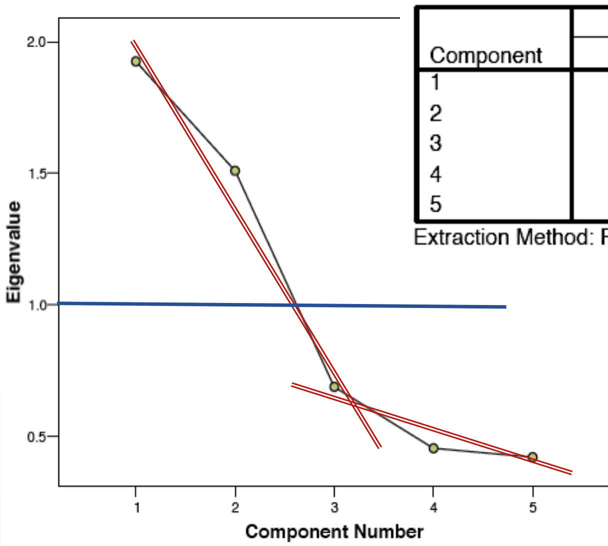
$|\mathbf{R} - \mathbf{I}\ \lambda_j| \rightarrow \mathbf{0}$

Redone one $\lambda$ after another,
always extracting the next largest one.

# Selection of Components




# Selection of Components



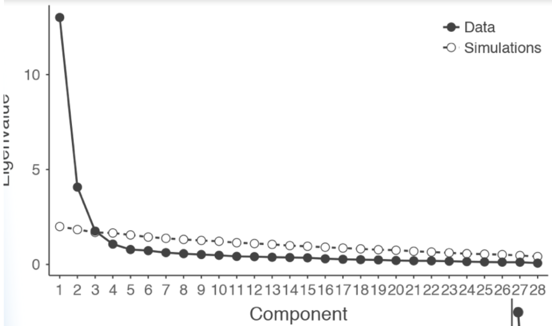

| Component | Initial Eigenvalues | | |
|---|---|---|---|
| | Total | % of Variance | Cumulative % |
| 1 | 1.925 | 38.505 | 38.505 |
| 2 | 1.510 | 30.207 | 68.712 |
| 3 | .688 | 13.761 | 82.473 |
| 4 | .455 | 9.102 | 91.575 |
| 5 | .421 | 8.425 | 100.000 |

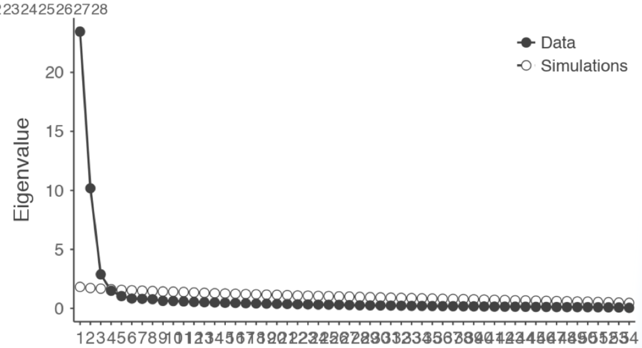
Extraction Method: Principal Component Analysis.

Criterion of lines crossing is not always precise

$\lambda > 1$ criterion overestimates PC number with large number of components

# Better: Simulate Chance




# Comparison