CSCI 491: Data Visualization

13- Dimension Reduction

Dimension Reduction

- When we have a very high-dimensional data set, we often want to lower the dimensionality of the problem.
- For example, we might have a corpus of unlabeled documents. Rather than using each word as a potential feature, we could instead develop a (shorter) list of topics and specify how related each document is to each topic.
- Similarly, even a collection of 100 x 100 pixel images would (naively) be a problem with 10,000 features. These features will be highly redundant, and we would instead like to build an ML model with a reduced set of independent features.
- Certain applications, such as reducing file sizes for faster data transfer, use dimensionality reduction.

Dimension Reduction

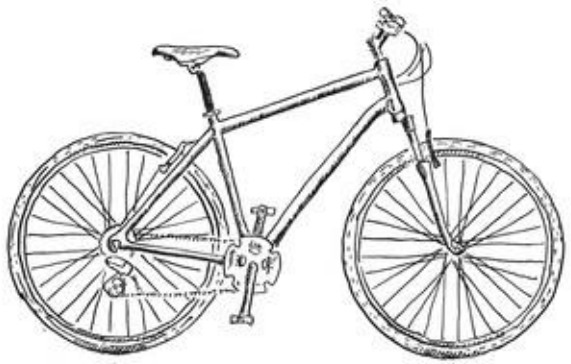
- The idea is to replace a large number of variables with a smaller number, which maintain a good representation of the data. In the map below, we can reduce/remove the landscape detail but still preserve the general geography.

Dimension Reduction

- What might be the intended use of a map like the one on the right? Is it as effective as the one on the left? Is it more effective?

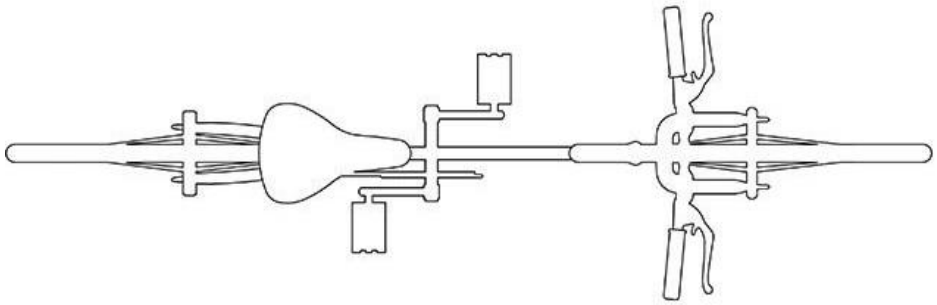
In-class activity

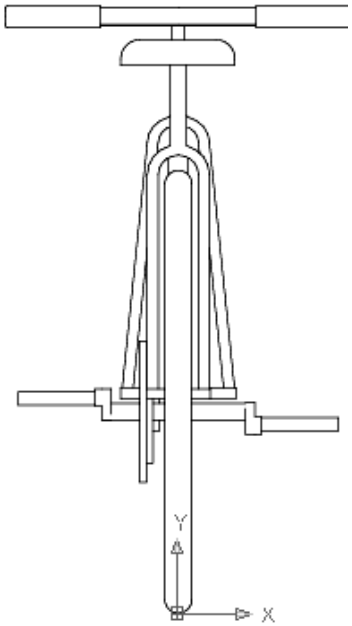
- Pick up a pen and paper and draw a simple diagram of a bicycle in 2 dimensions.



In-class activity

- Why not?





Dimension Reduction

- When we draw a bicycle diagram, we're projecting a 3D object onto a 2D plane. The reason most people's pictures of a bicycle end up looking the same is that we're drawing a bicycle along its most elongated axes. Such a projection captures a lot more information than drawing a bicycle from the top or front.

principal component analysis (PCA)

- This same concept is seen in principal component analysis (PCA), a mathematical technique which allows us to find a desirable projection that reduces the number of features in a matrix while preserving as much information as possible.
- The way PCA finds the projection that still maintains a good representation of the original data is by finding the directions in feature space along which the data has the largest spread or variability.

Principal Components Analysis

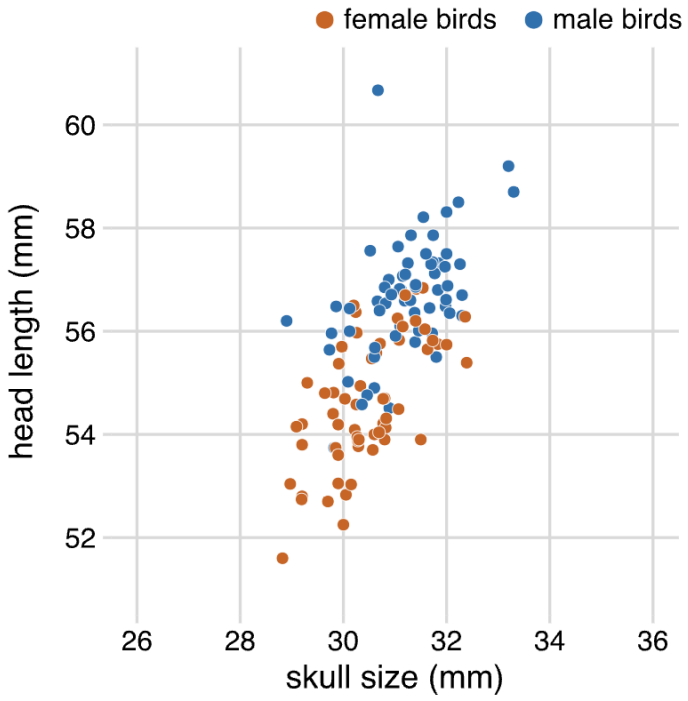
- PCA produces a low-dimensional representation of a dataset.
- It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- **Applications:**
- Producing derived variables for use in supervised learning problems (Feature selection)
- **PCA also serves as a tool for data visualization which is the focus of this lecture**

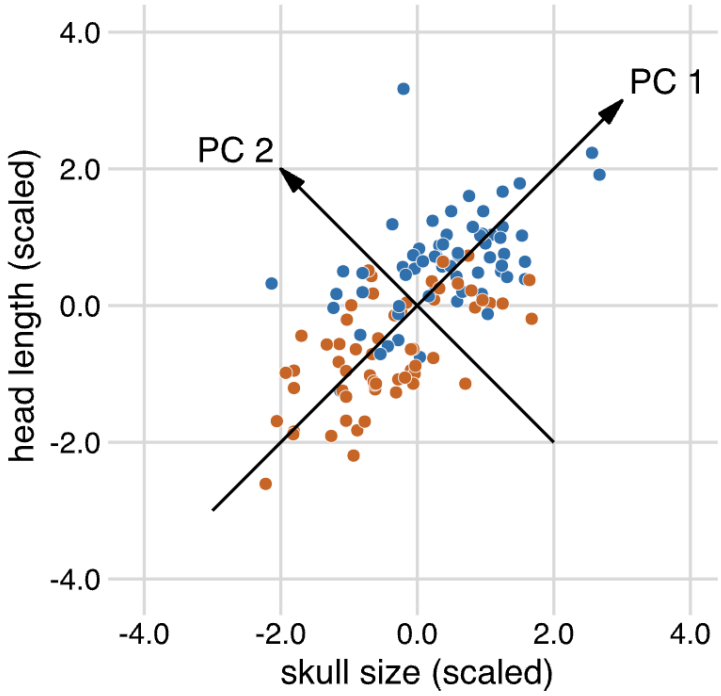
Principal Components Analysis

- Principal Components Analysis (PCA) is the most common dimension reduction approach to drive a low-dimensional set of features from a large set of variables.
- PCA is a technique for reducing the dimension of an n*p data matrix X to n*m and m less than or equal to p.
- The first principal component direction of the data is that along which the observations vary the most or the (normalized) linear combination of the variables with the largest variance.
- The second principal component has the largest variance, subject to being uncorrelated with the first and so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

PCA

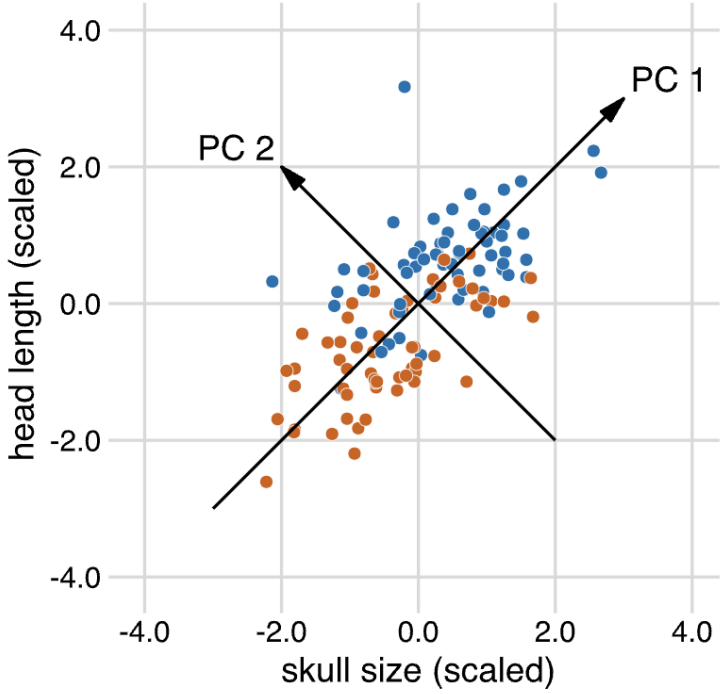
- Which direction got more variability?

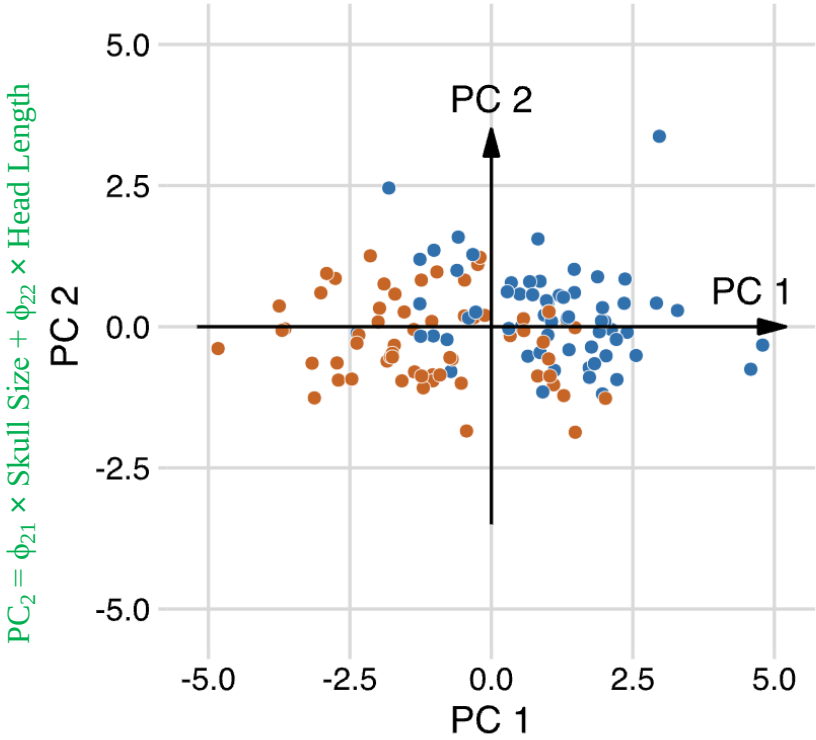




PCA is a rotation of the coordinate system

Skull size loading vector = (ϕ_{11} , ϕ_{21})
Head Length loading vector = (ϕ_{12} , ϕ_{22})





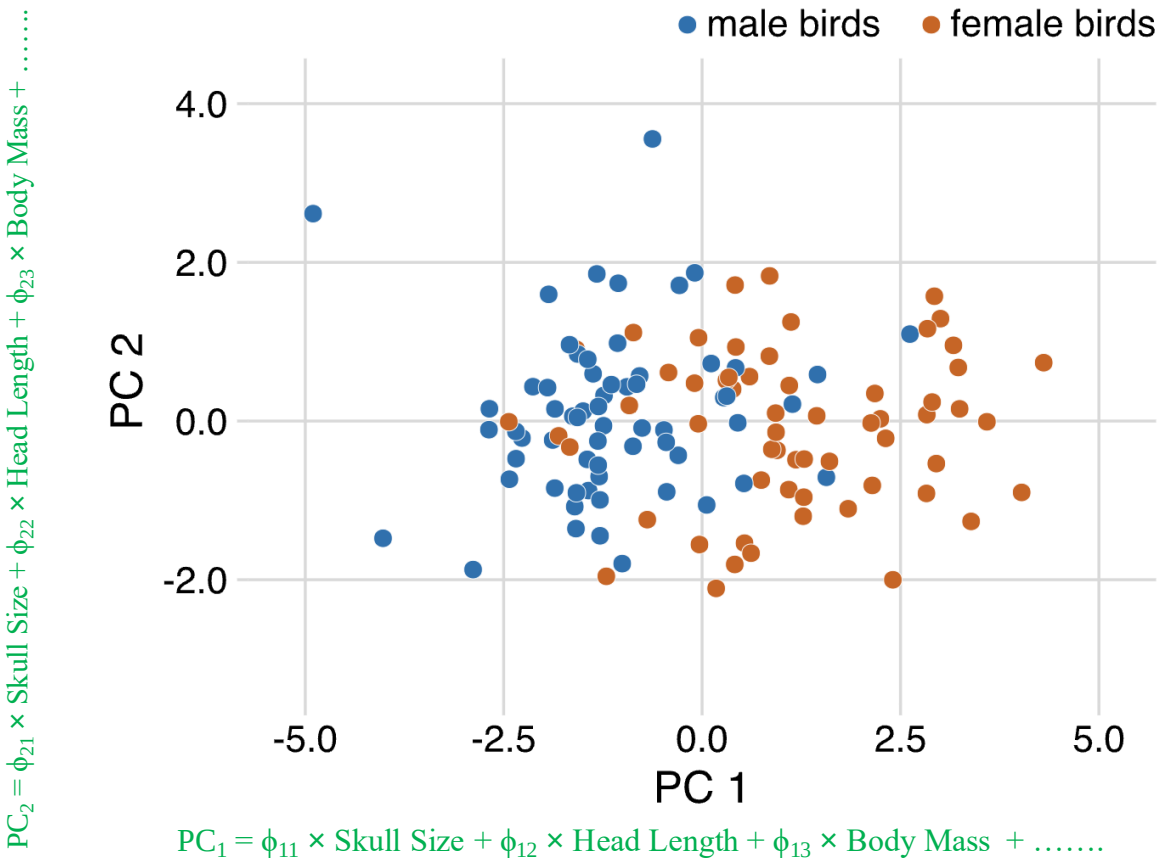
$PC_2 = \phi_{21} \times$ Skull Size + $\phi_{22} \times$ Head Length

$PC_1 = \phi_{11} \times$ Skull Size + $\phi_{12} \times$ Head Length

Blue Jays Dataset

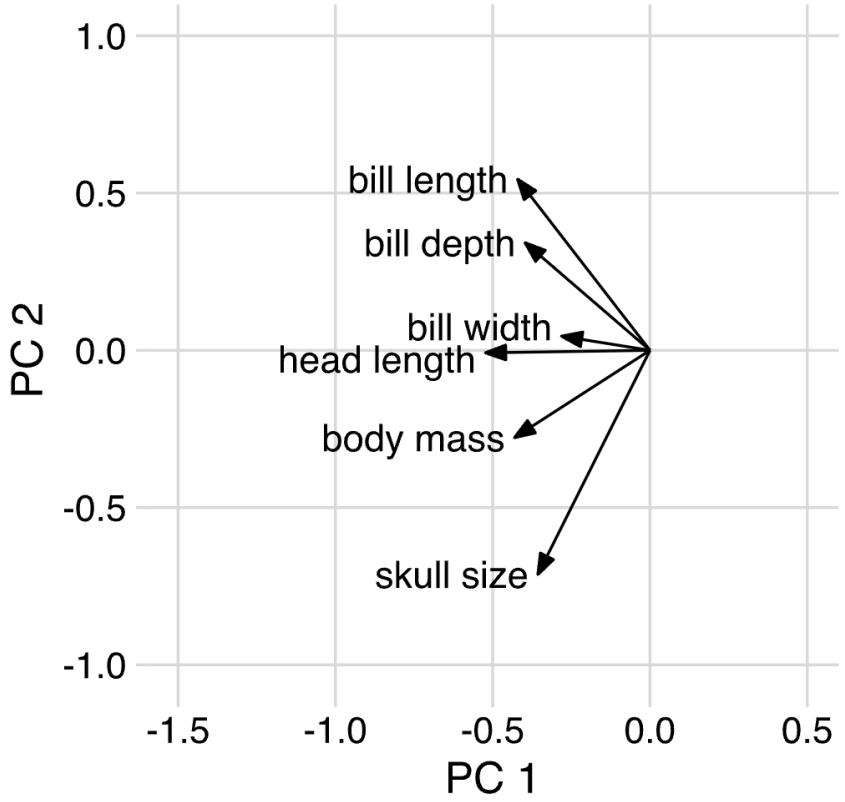
bird_id	sex	bill_depth_mm	bill_width_mm	bill_length_mm	head_length_mm	body_mass_g	skull_size_mm
0000-00000	M	8.26	9.21	25.92	56.58	73.30	30.66
1142-05901	M	8.54	8.76	24.99	56.36	75.10	31.38
1142-05905	M	8.39	8.78	26.07	57.32	70.25	31.25
1142-05907	F	7.78	9.30	23.48	53.77	65.50	30.29
1142-05909	M	8.71	9.84	25.47	57.32	74.90	31.85
1142-05911	F	7.28	9.30	22.25	52.25	63.90	30.00
1142-05912	M	8.74	9.28	25.35	57.12	75.10	31.77
1142-05914	M	8.72	9.94	30.00	60.67	78.10	30.67
1142-05917	F	8.20	9.01	22.78	52.83	64.00	30.05
1142-05920	F	7.67	9.31	24.61	54.94	67.33	30.33
1142-05930	M	8.78	8.83	25.72	56.54	76.40	30.82

PCA analysis of the entire blue jays dataset



Male and female birds separate along PC1

The rotation matrix allows us to interpret the PCs



Skull size loading vector = (ϕ_{11} , ϕ_{21})
Head Length loading vector = (ϕ_{12} , ϕ_{22})
Body Mass loading vector = (ϕ_{13} , ϕ_{23})

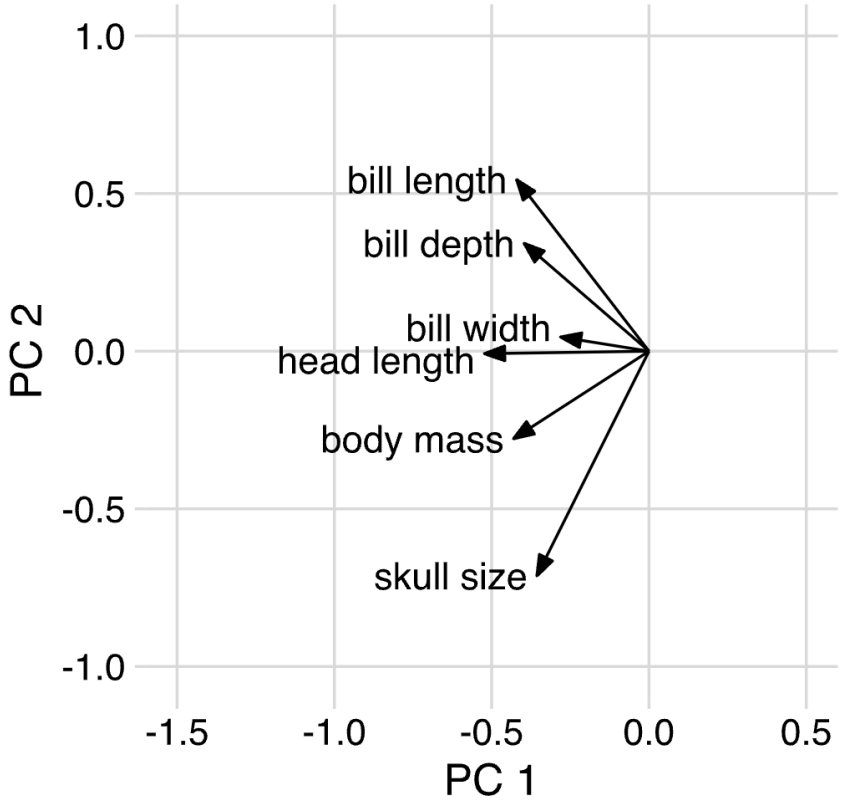
•
•
•

$PC_2 = \phi_{21} \times$ Skull Size $+ \phi_{22} \times$ Head Length $+ \phi_{23} \times$ Body Mass + …….

$PC_1 = \phi_{11} \times$ Skull Size $+ \phi_{12} \times$ Head Length $+ \phi_{13} \times$ Body Mass + …….

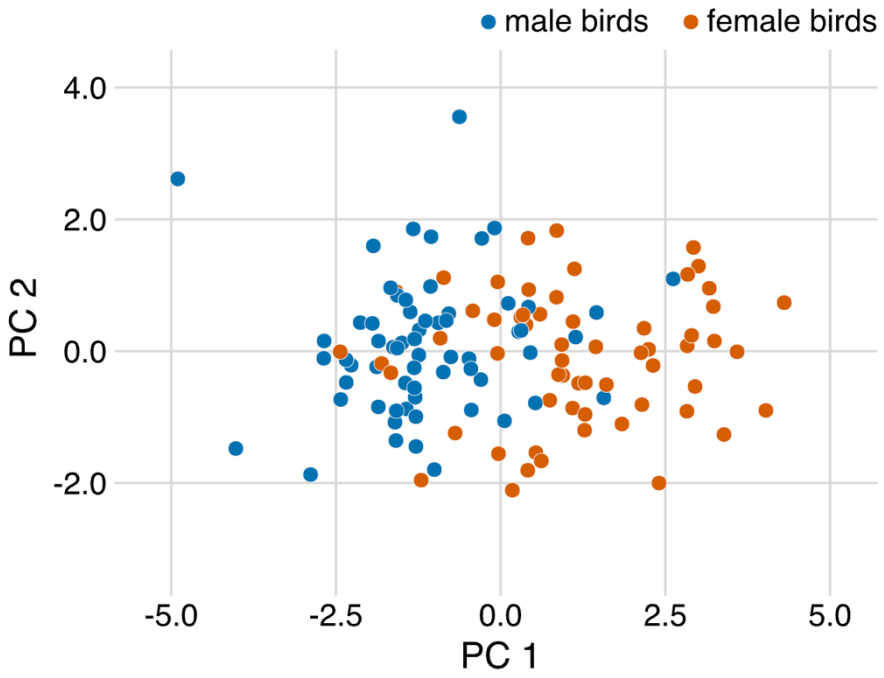
All variables contribute negatively to PC1; it represents the overall size of the bird

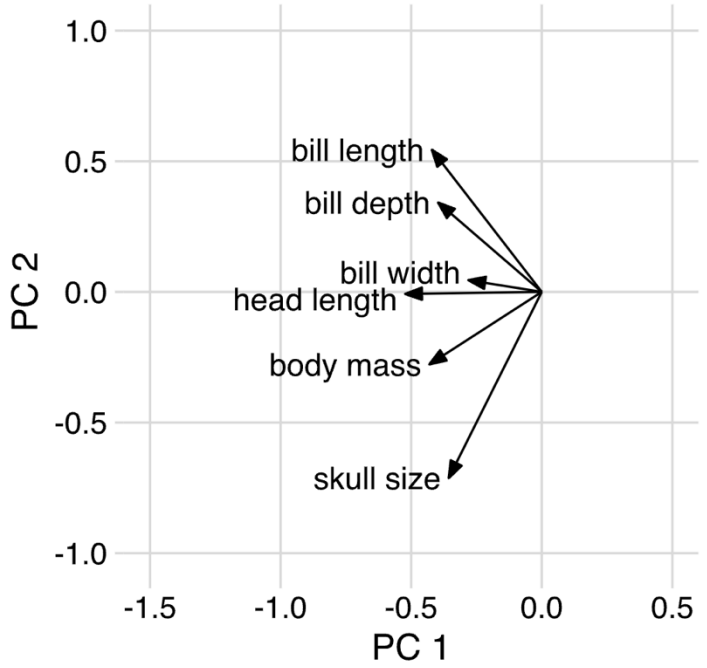
The rotation matrix allows us to interpret the PCs



PC2 represents the difference between bill length and skull size

The rotation matrix allows us to interpret the PCs





Male birds are larger than female birds

Both male and female birds have long and short bills relative to their overall size

US Arrests Example

- USAarrests data: For each of the fifty states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. We also record UrbanPop (the percent of the population in each state living in urban areas).
- The principal component score vectors have length n = 50, and the principal component loading vectors have length p = 4.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

US Arrests: PCA plot

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

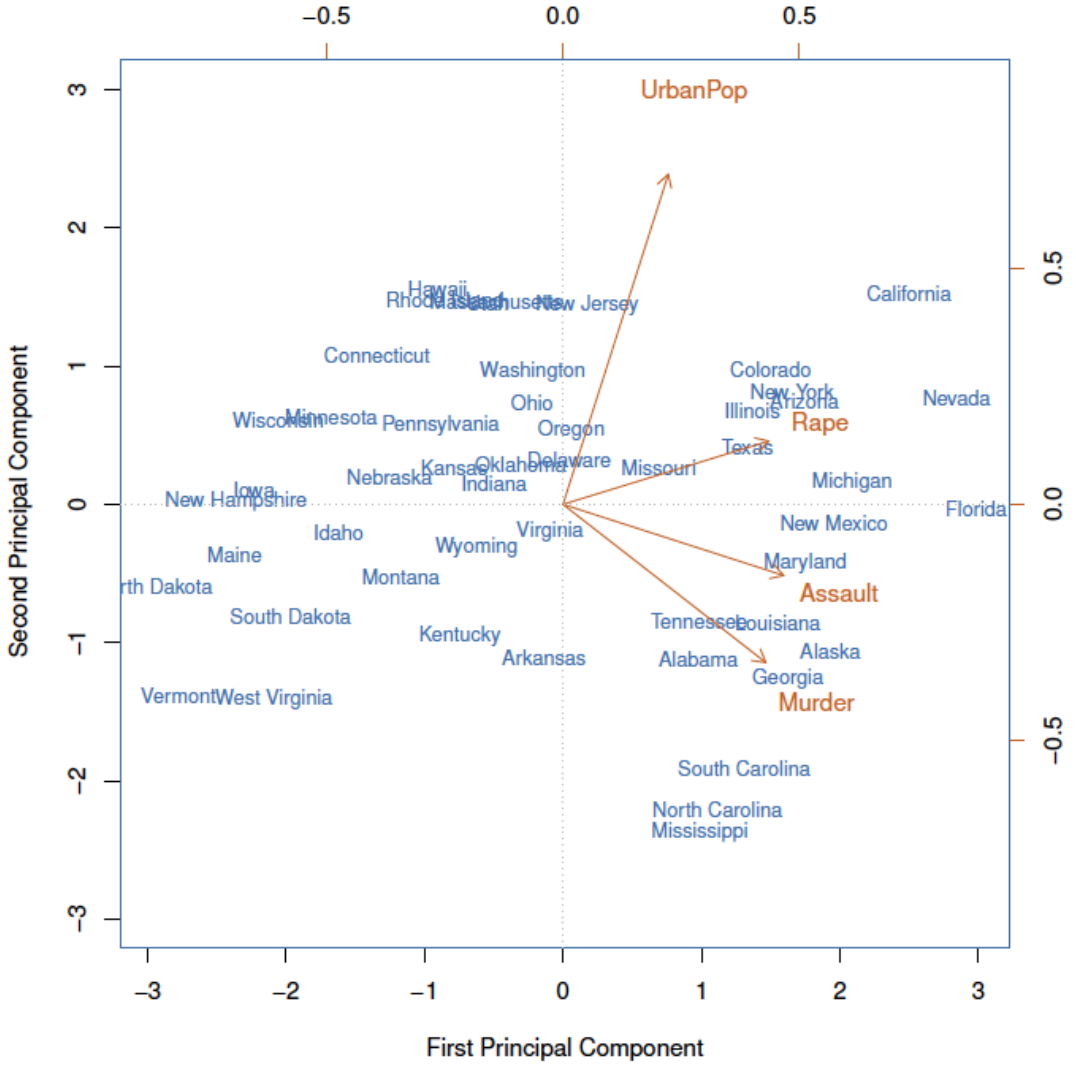
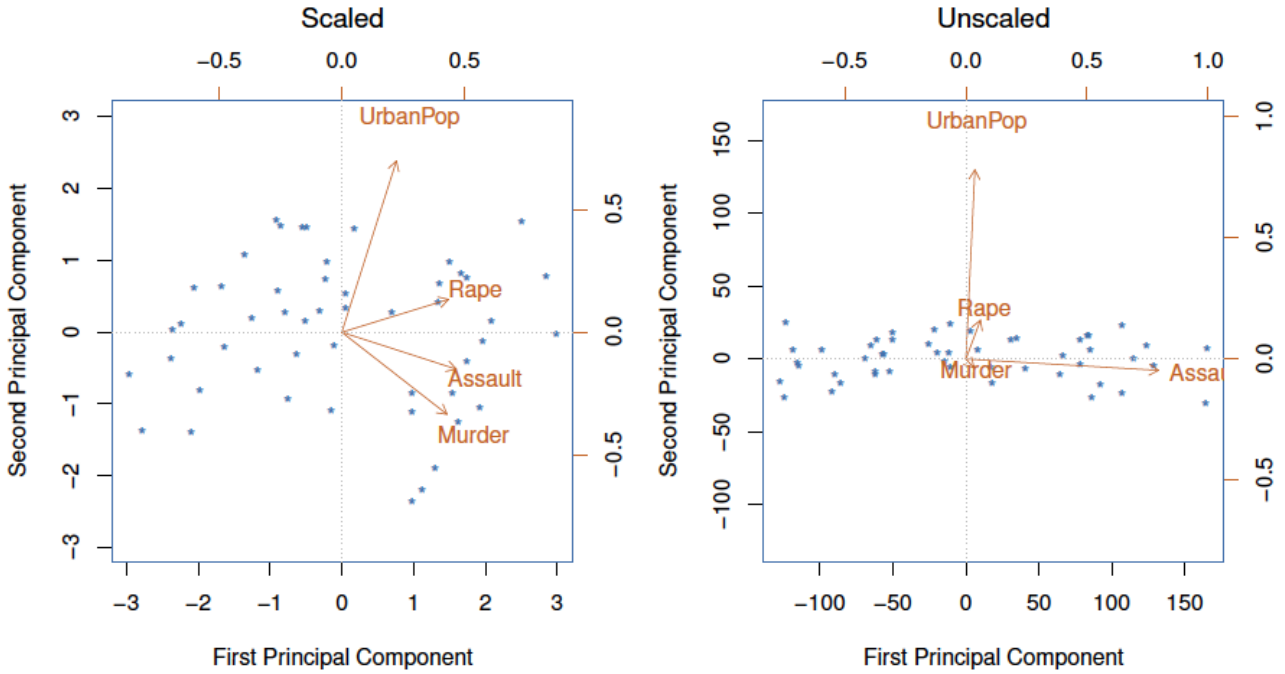


Figure details

- The first two principal components for the USArrests data. The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right). For example, the loading for Rape on the first component is 0.54, and its loading on the second principal component 0.17 [the word Rape is centered at the point (0.54, 0.17)].
- This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.

Scaling of the variables matters

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2,$$

and the variance explained by the mth principal component is

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{} z_{im}^2.$$

It can be shown that

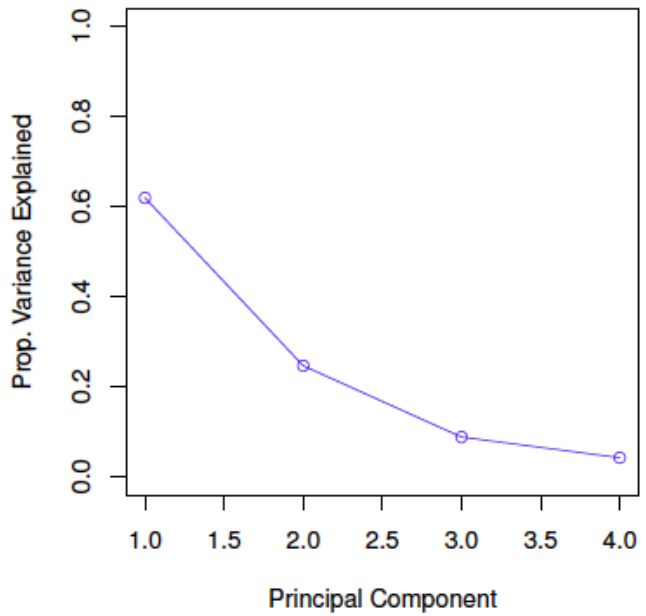
$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{m=1}^{M} \mathrm{Var}(Z_m),$$

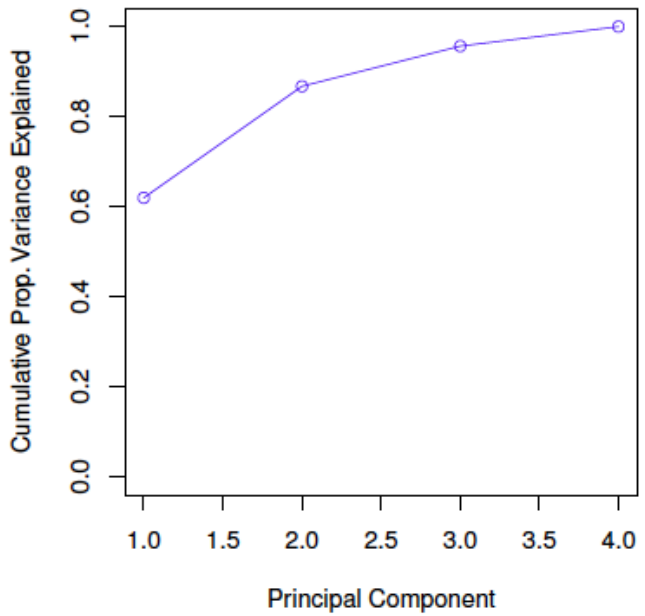
with M = min(n-1, p).

Proportion Variance Explained

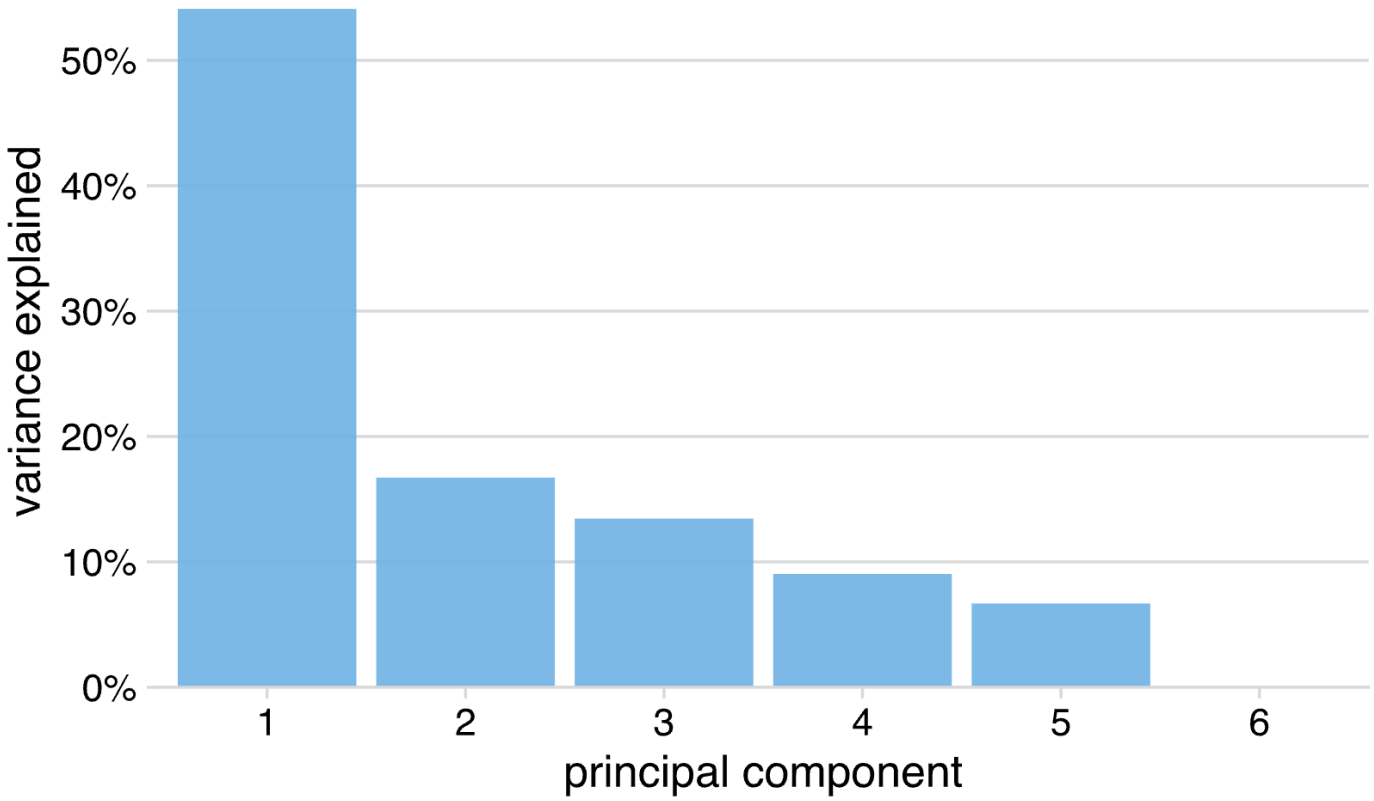
- Therefore, the PVE of the mth principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}.$$





Proportion Variance Explained for Blue Jays



PC 1 captures over 50% of the total variance in the dataset

Overall bird size explains >50% of the variation in the various measurements