Unsupervised Learning

Nayel Bettache

Unsupervised Learning

- In the supervised learning setting, we typically have access to a set of p features $X_1, X_2, \ldots, X_p$, measured on n observations, and a response Y also measured on those same n observations. The goal is then to predict Y using $X_1, X_2, \ldots, X_p$.
- In unsupervised learning, we have only a set of features $X_1, X_2, \ldots, X_p$ measured on n observations. We are not interested in prediction, because we do not have an associated response variable Y.
- The goal is to discover interesting patterns about the measurements. (e.g., visualization, find subgroup of data, find subgroup of features, find independence among features)

Two important unsupervised learning problems

- **Principal components analysis**, a tool used for data visualization or data pre-processing before supervised techniques are applied.
- **Clustering**, a broad class of methods for discovering unknown subgroups in data.

Principal Components Analysis

- PCA allows us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set.
- 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.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principal Components Analysis

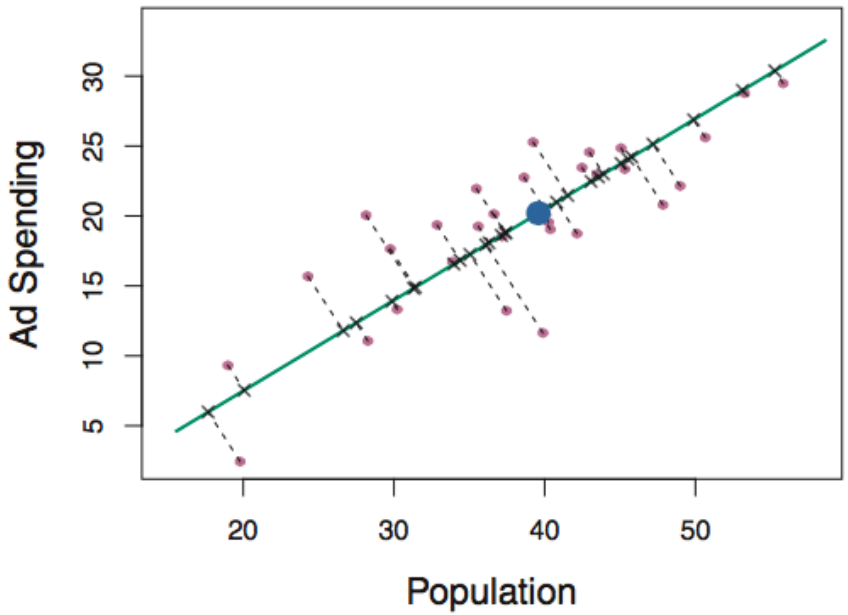
- How should we visualize n observations with p features X_1,X_2,...,X_p?
- The first principal component of a set of features X_1,X_2,...,X_p is the normalized linear combination of the features

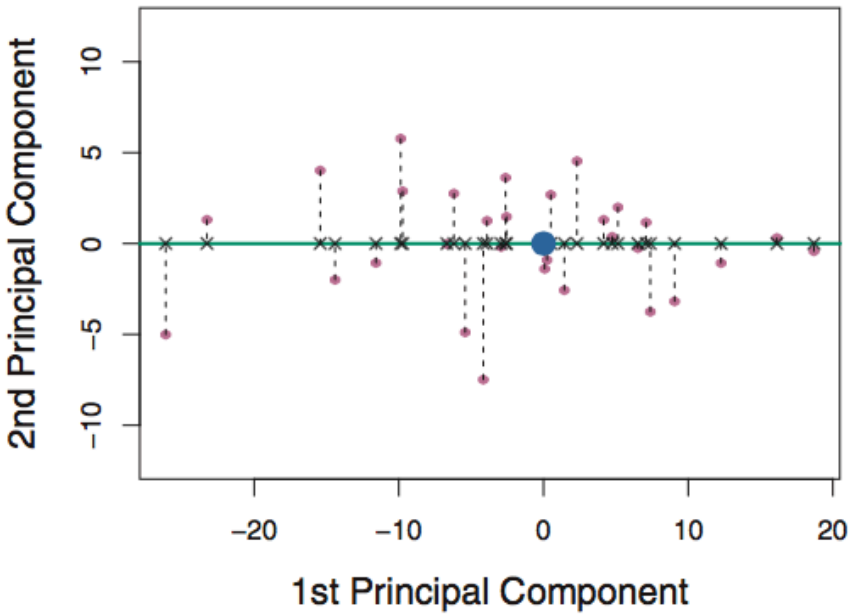
$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p$$

  that has the largest variance. By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- We refer to the elements $\phi_{11}, ..., \phi_{p1}$ as the **loadings** of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \phi_{21}, ..., \phi_{p1})^T$.
- We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

Advertising Data

Consider two features: population size (pop) and ad spending for a particular company (ad).





The first principal component direction is shown in green. It is the dimension along which the data vary the most. We get

$$Z_1 = 0.839 \times (pop - \bar{pop}) + 0.544 \times (ad - \bar{ad}).$$

Computation of Principal Components

- Suppose we have a $n \times p$ data set $\mathbf{X}$. Since we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero (that is, the column means of $\mathbf{X}$ are zero).
- We then look for the linear combination of the sample feature values of the form

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + ... + \phi_{p1} x_{ip},$$

  for $i = 1, ..., n$ that has the largest variance, subject to $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- Since each of the x_{ij} has mean zero, then so does z_{i1} (for any values of ϕ_{j1}). Hence the sample variance of the z_{i1} can be written as $n^{-1} \sum_{i=1}^{n} z_{i1}^2$.

Computation of Principal Components

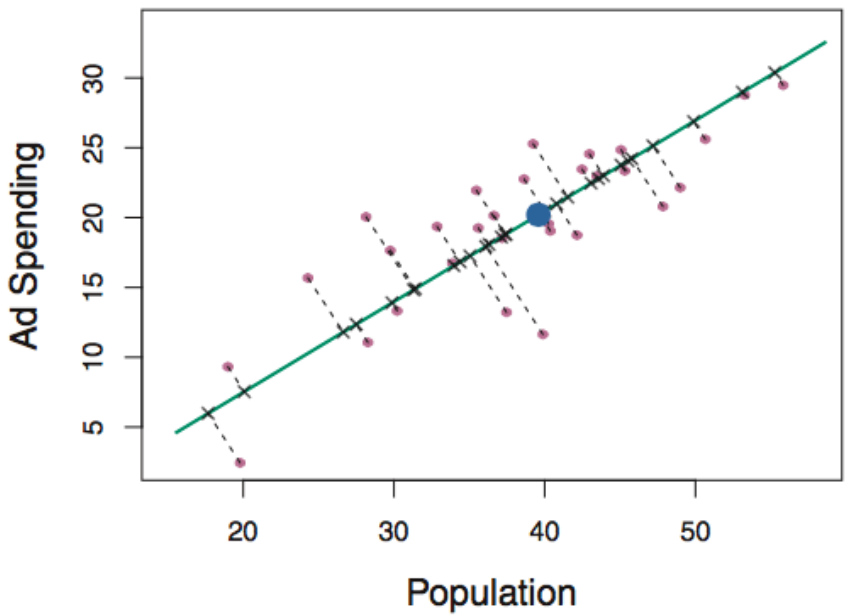
- The first principal component loading vector solves the optimization problem

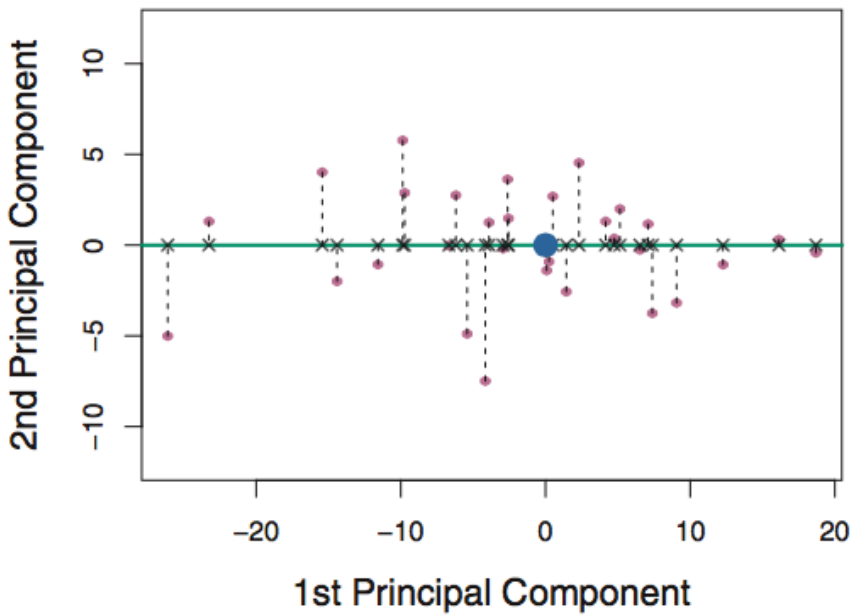
$$\max_{\phi_1} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j1} x_{ij})^2 \right\}, \quad st \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

- We refer to $z_{11}, ..., z_{n1}$ as the **scores** of the first principal component.
- The problem can be solved via an eigen decomposition, a standard technique in linear algebra.

Geometry of PCA

- The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}, ..., \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, ..., x_n$ onto this direction, the projected values are the principal component scores $z_{11}, ..., z_{n1}$ themselves.





$$z_{i1} = 0.839 \times (pop_i - \bar{pop}) + 0.544 \times (ad_i - \bar{ad}).$$

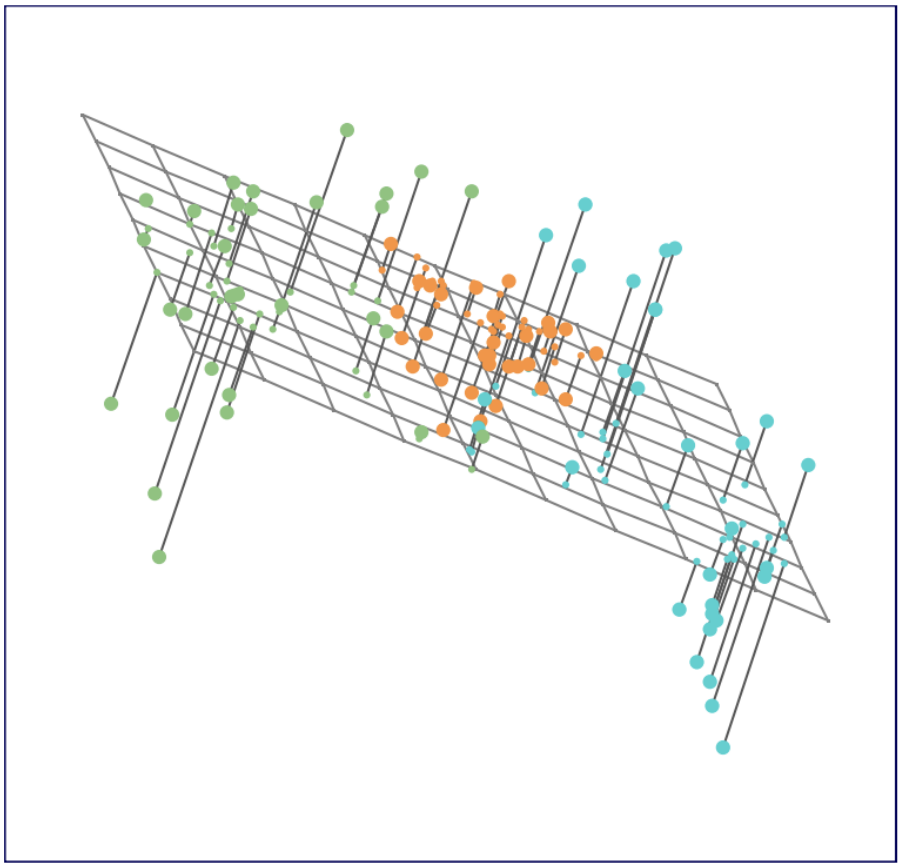
Second principal component

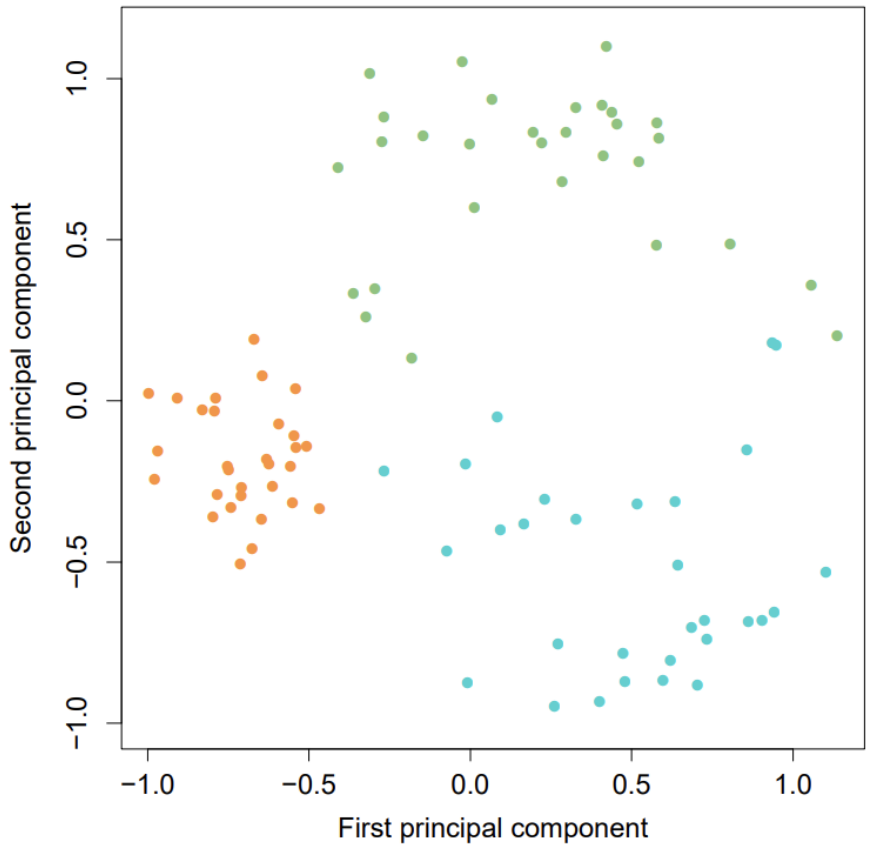
- The second principal component is the linear combination of $X_1, ..., X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- The second principal component scores $z_{12}, z_{22}, ..., z_{n2}$ take the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + ... + \phi_{p2}x_{ip},$$

  where ϕ_2 is the second principal component loading vector, with elements $\phi_{12}, ..., \phi_{p2}$.
- It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1.

Visualization

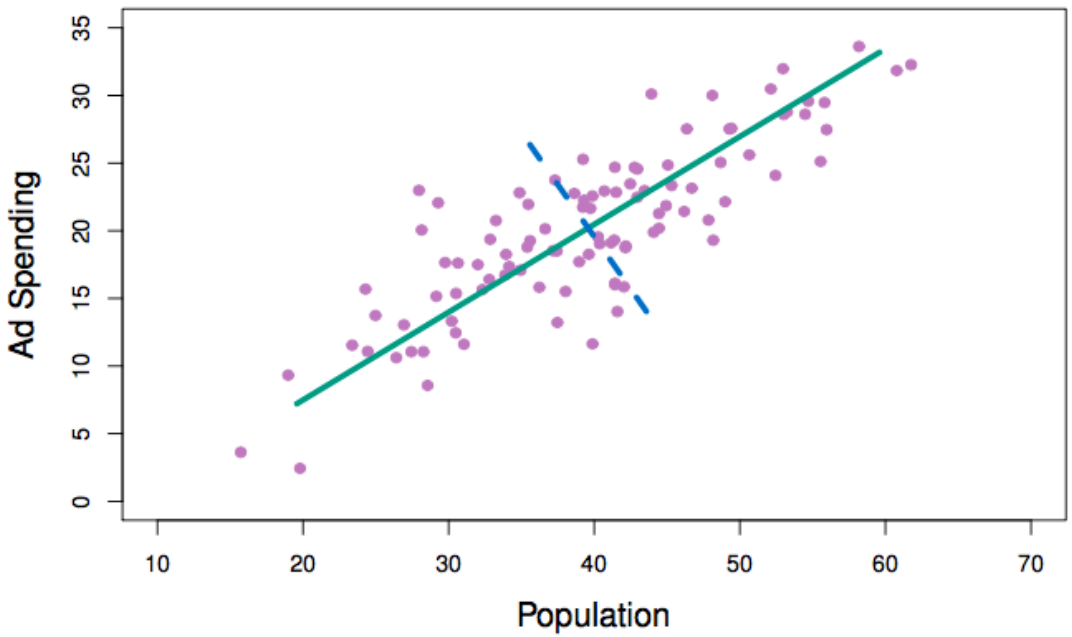




- 90 observations simulated in three dimensions.
- Left: the first two principal component directions span the plane that best fts the data. The plane is positioned to minimize the sum of squared distances to each point.
- Right: the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane.

Advertising Data

Consider two features: population size (pop) and ad spending for a particular company (ad).

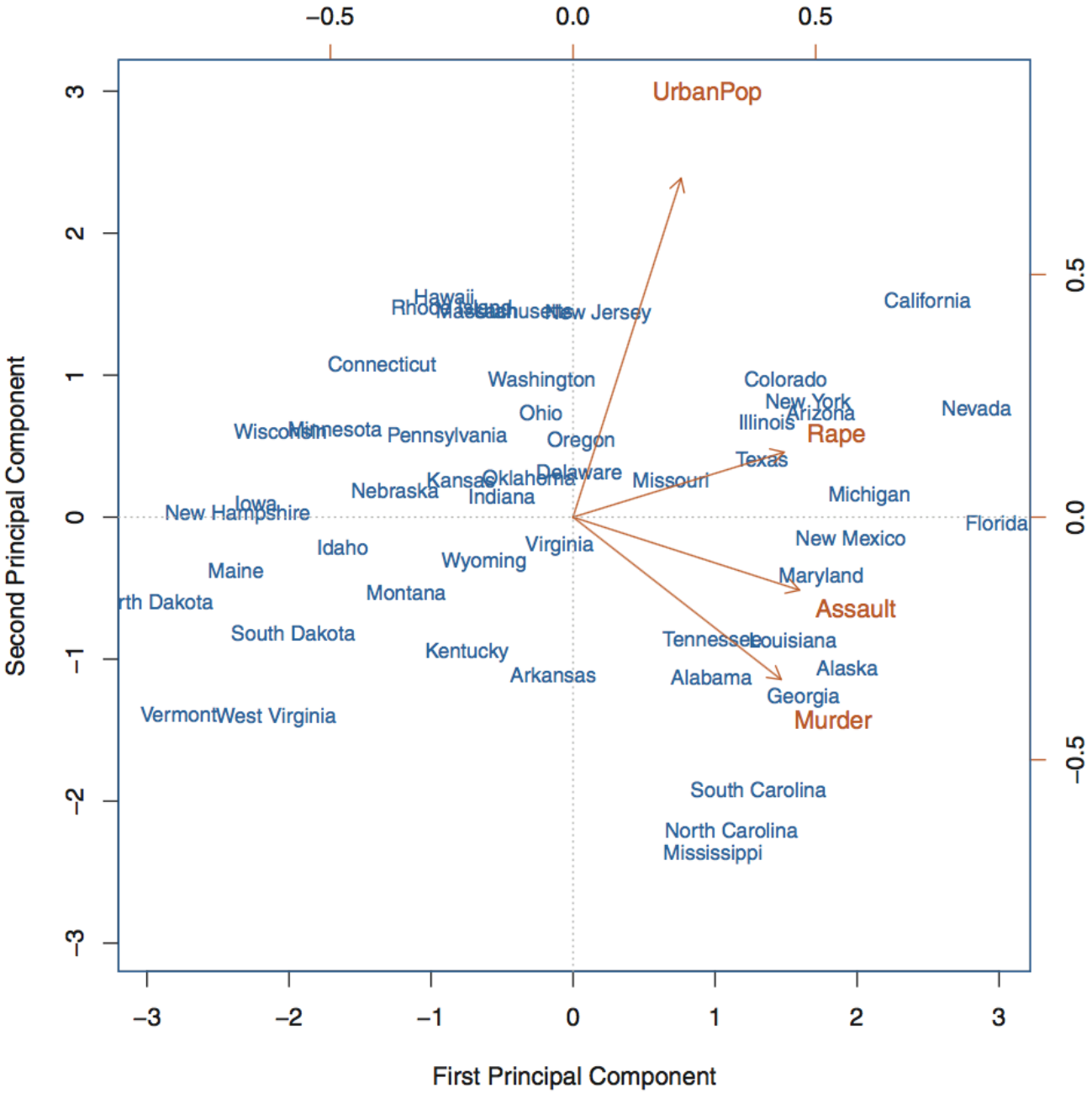


The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

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.

USAarrests data



USAarrests data

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.

USAarrests data

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

The principal component loading vectors, ϕ_1 and ϕ_2, for the USArrests data.

Some practical considerations

- In general, scaling the variables to have standard deviation one is recommended.
- Each principal component loading vector is unique, up to a sign flip.
- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.

Proportion Variance Explained

- The total variance present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} var(X_j) \approx \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2.$$

  the variance explained by the mth principal component is

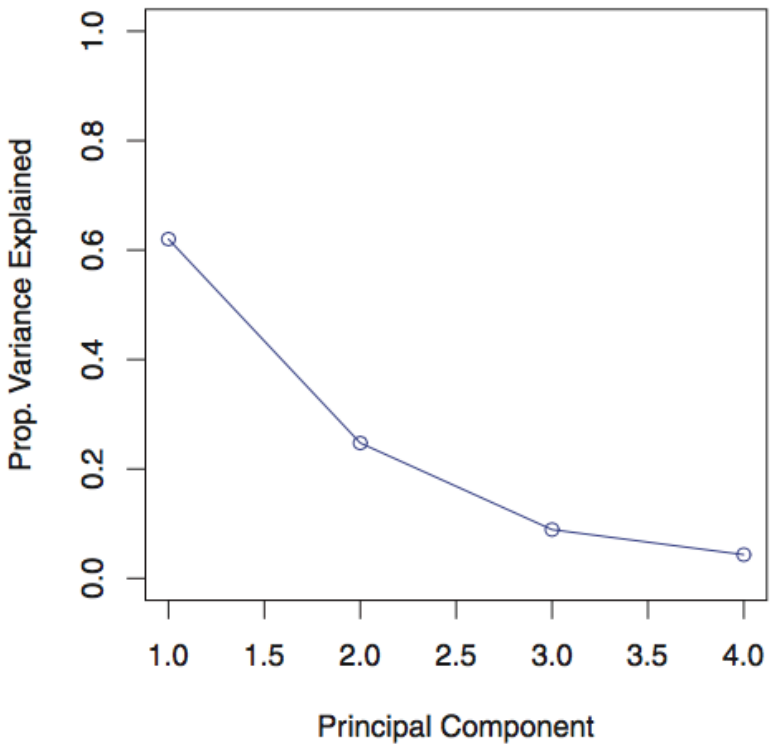
$$var(Z_m) \approx \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

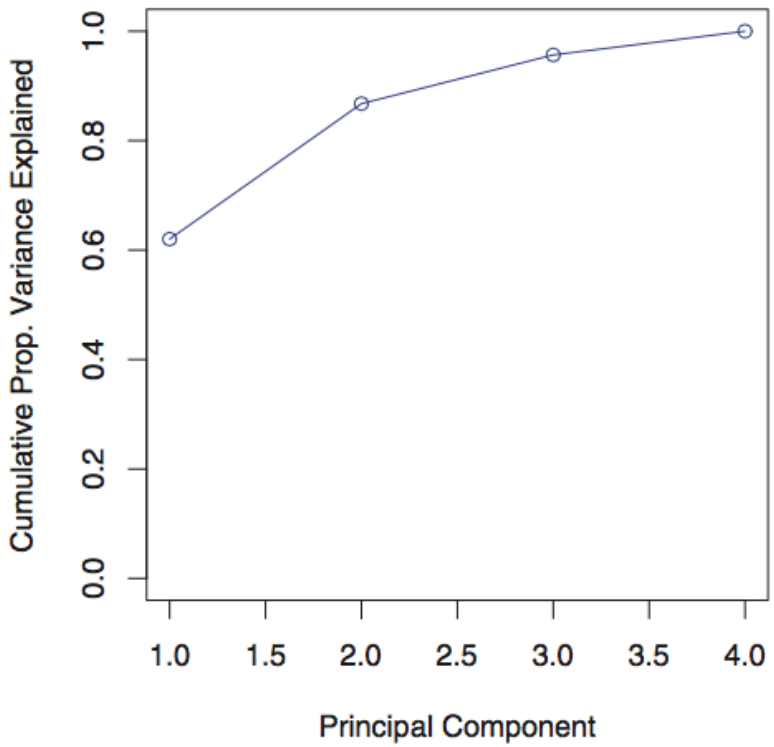
- The PVE of the mth principal component is given by the positive quantity between 0 and 1

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

- The PVEs sum to one. We sometimes display the cumulative PVEs.

USAarrests data





Deciding How Many Principal Components to Use

- In general, a $n \times p$ data matrix $\mathbf{X}$ has $\min(n-1, p)$ distinct principal components.
- We would like to use the smallest number of principal components required to get a good understanding of the data. How many principal components are needed?
- No simple answer to this question, as cross-validation is not available for this purpose. (It is possible in PCR!)
- An adhoc approach is to look at the PVE plot.