PRINCIPAL COMPONENT ANALYSIS

INTRODUCTION

- Principal components allow to summarize a large set of correlated variables with a small number of uncorrelated variables that collectively explain most of the variability in the original set
- PCA 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
- PCA refers to the process by which principal components are computed, and the subsequent use of these components in data understanding

PCA

- The first principal component of a set of features

$$X_1, X_2, \dots, X_p$$

is the normalized linear combination

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

that has the largest variance.

- By normalized, we mean that

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

PCA

- Elements $\phi_{11}, \dots, \phi_{p1}$ are known as loadings of the first component.
- Vector $\phi_1 = (\phi_{11}, \dots, \phi_{p1})$ is the principal component loading vector.
- Then, we look for the linear combination

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{p1}x_{ip}, i = 1, \dots, n$$

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

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$$

PCA

- Loading vector can be determined from the following optimization problem

$$\max_{\phi_{11},\ldots,\phi_{p1}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^{2}$$

$$\text{subject to} \quad \sum_{j=1}^{p}\phi_{j1}^{2} = 1$$

PCA

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

PCA: FURTHER COMPONENTS

- The second principal component Z_2 is the linear combination of $X_1, \dots, X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- The second principal component scores $z_{12}, z_{22}, \dots, z_{n2}$ take the form

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}, \qquad \sum_{j=1}^{p} \phi_{j2}^2 = 1$$

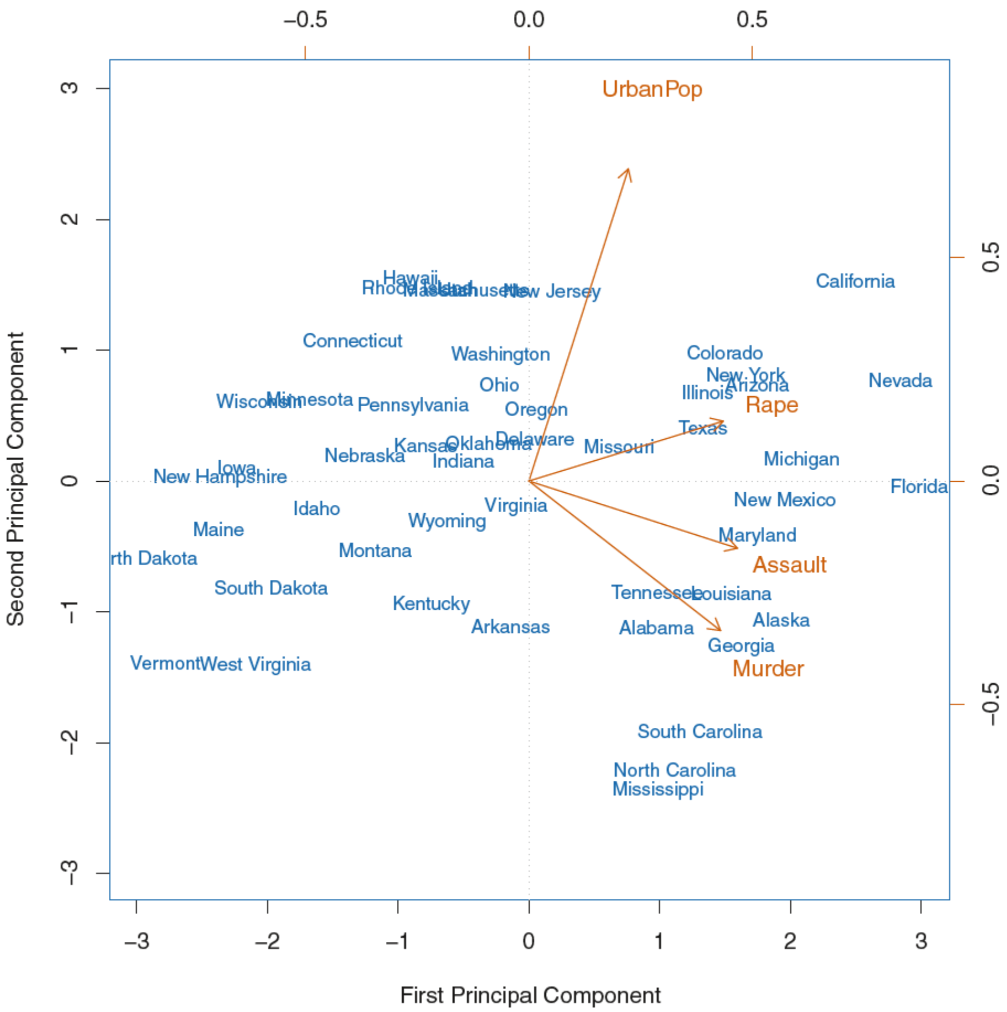
- $\phi_2 = (\phi_{12}, \dots, \phi_{p2})$ is the second principal component loading vector.
- It turns out that constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction ϕ_2 to be orthogonal (perpendicular) to the direction ϕ_1.

PCA: FURTHER COMPONENTS

- The third principal component Z_3 is the linear combination of $X_1, \dots, X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1 and Z_2.
- And so on until we construct the all needed components

$$Z_1, \dots, Z_m$$

USARRESTS DATA: A BIPLOT



USARRESTS DATA: A BIPLOT

- This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.
- 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).

USARRESTS DATA: A BIPLOT

- The first loading vector places approximately equal weight on Assault, Murder, and Rape, with much less weight on UrbanPop.
- Hence this component roughly corresponds to a measure of overall rates of serious crimes.
- The second loading vector places most of its weight on UrbanPop and much less weight on the other three features.
- Hence, this component roughly corresponds to the level of urbanization of the state.

USARRESTS DATA: A BIPLOT

- Overall, we see that the crime-related variables (Murder, Assault, and Rape) are located close to each other, and that the UrbanPop variable is far from the other three.
- This indicates that the crime-related variables are correlated with each other—states with high murder rates tend to have high assault and rape rates—and that the UrbanPop variable is less correlated with the other three.

USARRESTS DATA: A BIPLOT

- The states with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates
- The state like North Dakota, with negative scores on the first component, have low crime rates.
- California also has a high score on the second component, indicating a high level of urbanization, while the opposite is true for states like Mississippi.
- States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

PROPORTION OF VARIANCE EXPLAINED

- Total variance of the original space

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

- The variance explained by the m-th principal component

$$\frac{1}{n} \sum_{j=1}^{p} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

- PVE of the m-th component

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

SINGULAR VALUE DECOMPOSITION (SVD)

- Formally, the singular value decomposition (SVD) of an $n \times p$ matrix X is a factorization of the form

$$X = U\Sigma V^T$$

- U is an $n \times n$ unitary matrix,
- Σ is a $n \times p$ rectangular diagonal matrix with non-negative real numbers on the diagonal,
- V is an $p \times p$ unitary matrix.

SVD

- The diagonal entries σ_i are known as the singular values of X. Non-zero singular values are the square roots of the non-zero eigenvalues of both $X^T X$.
- The columns of U and the columns of V are called the left-singular vectors and right-singular vectors of M, respectively. The right-singular vectors of X are a set of orthonormal eigenvectors of $X^T X$.
- The principal component directions

$$\phi_1 = (\phi_{11}, \dots, \phi_{p1}), \phi_2 = (\phi_{12}, \dots, \phi_{p2}), \phi_3 = (\phi_{13}, \dots, \phi_{p3}), \dots$$

are the ordered sequence of right singular vectors of the matrix X, and the variances of the components are $1/n$ times the squares of the singular values.

CORRELATION MATRIX APPROACH

- Step 1: Calculate the correlation matrix C ($p \times p$ matrix).
- Step 2: Calculate the eignevalues $\lambda_1, \dots, \lambda_p$ of the correlation matrix.
- Step 3: Calculate the corresponding normalized eigenvectors $v_1, \dots, v_p$ which are orthogonal to each other and compose the principal directions.
- Step 4: $Z = X\,V$, where $V = [v_1, \dots, v_p]$.

CORRELATION MATRIX APPROACH

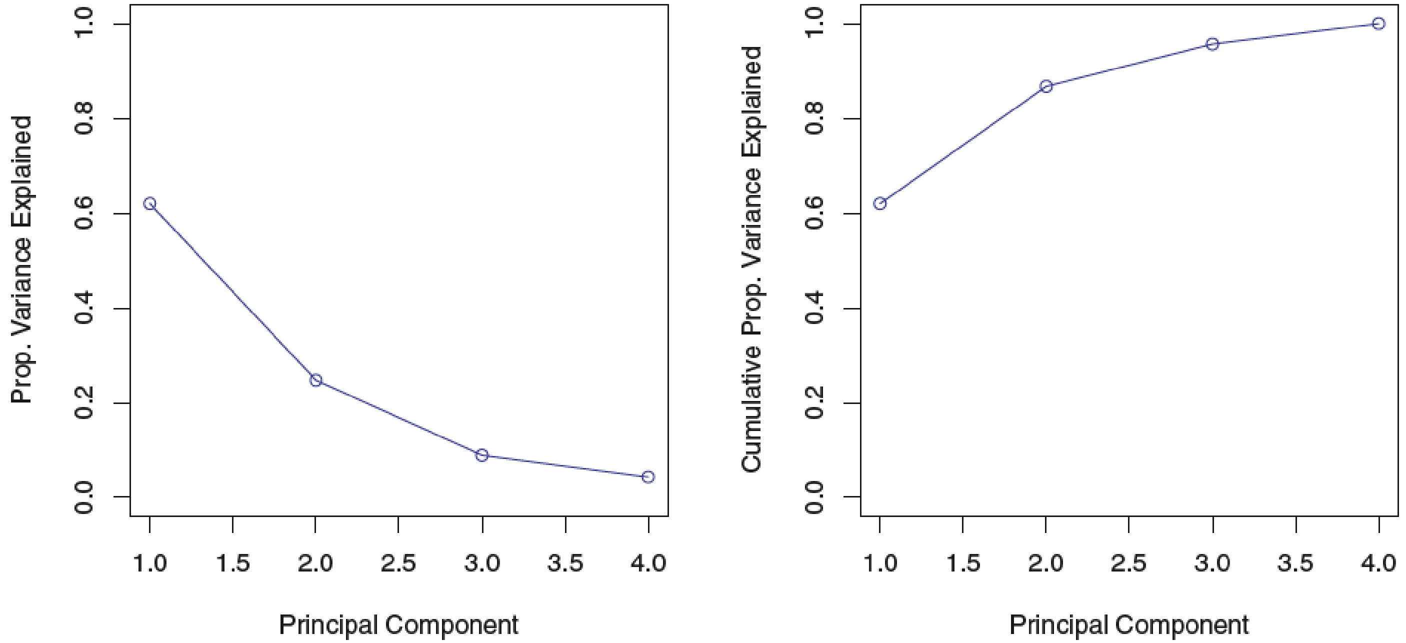
How many principal components we need?

- Step 1: Take a parameter $\varepsilon > 0$. Usually, $\varepsilon = 0.8 \sim 0.9$
- Step 2: If

$$\frac{\lambda_1 + \cdots + \lambda_{k-1}}{\lambda_1 + \cdots + \lambda_p} < \varepsilon, \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \geq \varepsilon$$

- Then k is the needed number of the principal components for that specific application.

A SCREE PLOT



- **Left:** a scree plot depicting the proportion of variance explained by each of the four principal components
- **Right:** the cumulative proportion of variance explained by the four principal components

HOW MANY PRINCIPAL COMPONENTS?

If we use principal components as a summary of our data, how many components are sufficient?

- No simple answer to this question in case of unsupervised learning
- In case of supervised learning, a cross-validation can help to assess the prediction accuracy for a given number of principal components

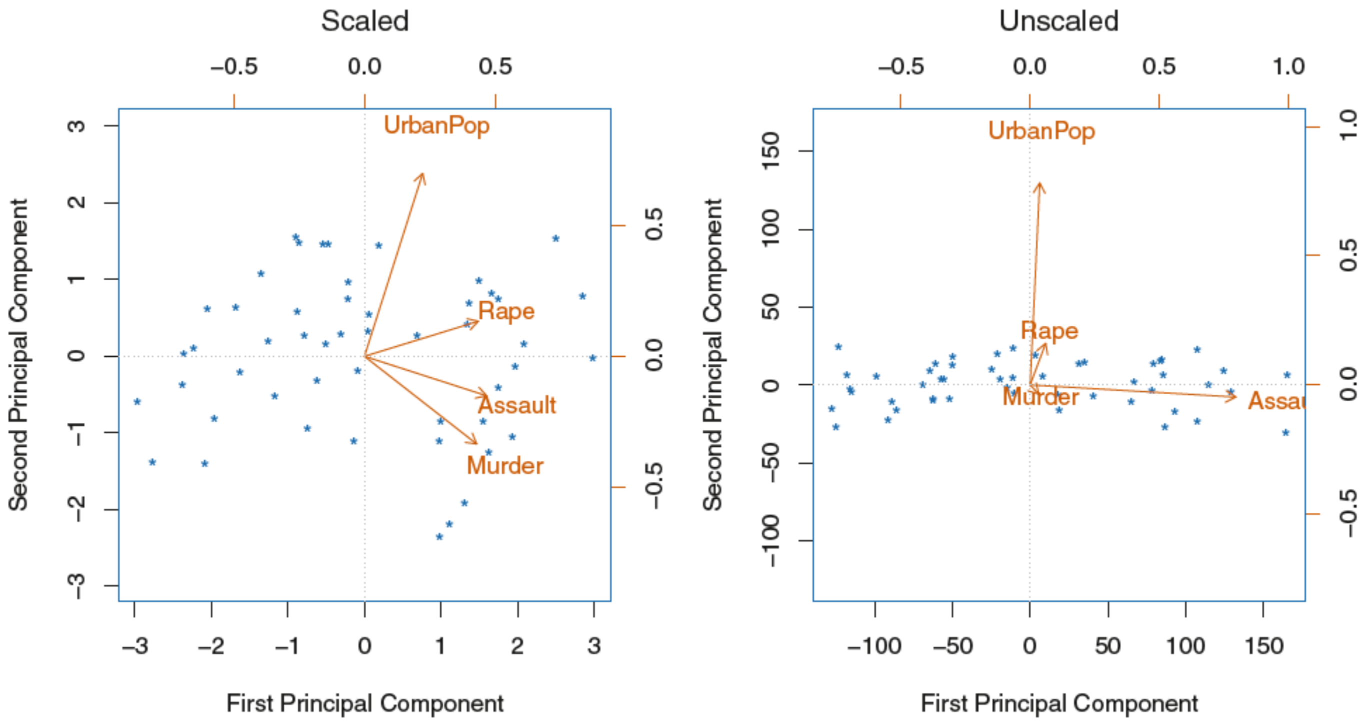
APPLICATIONS

- Dimensionality Reduction
- Data Visualization
- Data Compression

SCALING OF THE VARIABLES

- 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.

SCALING OF THE VARIABLES



- Left: The variables scaled to have unit standard deviations.
- Right: principal components using unscaled data. Assault has by far the largest loading on the first principal component because it has the highest variance among the four variables.