MATH 189
Principal Component Analysis

Wenxin Zhou
UC San Diego

Time: 2:00—3:20 & 3:30—4:50pm TueThur
Location: CENTR 115

Outline

- In the last two lectures, we introduced the method of least squares for linear regression and maximum likelihood estimation for logistic regression.
  - Least squared estimator for linear regression: close-form, test marginal effect, test model adequacy
  - Maximum likelihood estimator for logistic regression: no closed-form, test marginal effect, generalized likelihood ratio test
- Today we will introduce a dimension reduction technique, named principal component analysis.
  - High-dimensional data and dimension reduction
  - Principal component analysis

High-Dimensional Dataset

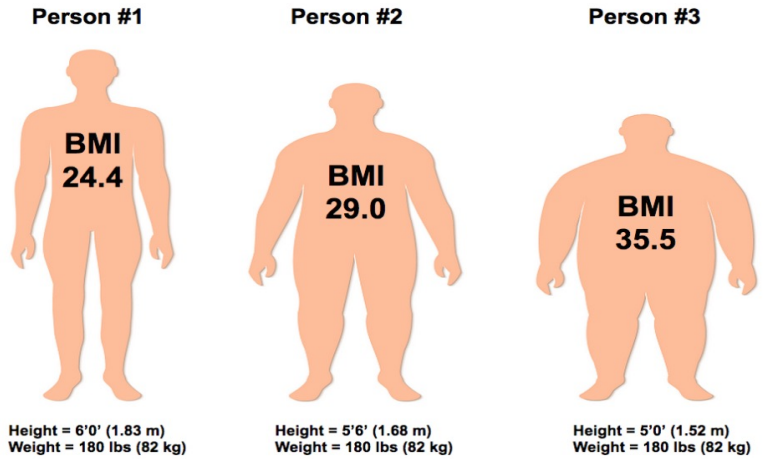
- Most classical statistical techniques for regression and classification are intended for the low-dimensional setting in which n, the number of observations, is much larger than p, the number of features.
- In the early stage of many scientific fields, the collected dataset is multivariate but low-dimensional in nature. Also we have technical difficulty to collect, store and compute huge datasets.

- In the past two decades, new technologies have changed the way in which data are collected in fields ranging from medicine, finance, to marketing and social network.
- It is now commonplace to collect as many features as possible (p very large). While p is growing fast, the number of observations n is often limited due to cost, sample availability, among other considerations.

Example: Statistical Model for Obesity

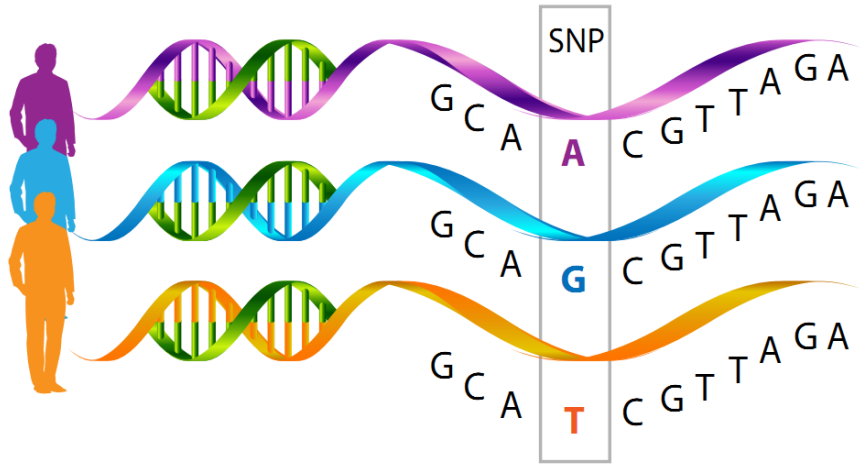
Obesity Model in 1970

- The classic BMI (body mass index) model was estimated on the base of age, gender, height and weight data collected from a group of people, say 200.
- $n = 200$ and $p = 4$.



Obesity Model in 2010

- Modern obesity model also collects measurements for half a million of genetic variations from people for inclusion in the predictive model.
- $n = 200$ and $p \approx 500{,}000$.



Dimension Reduction

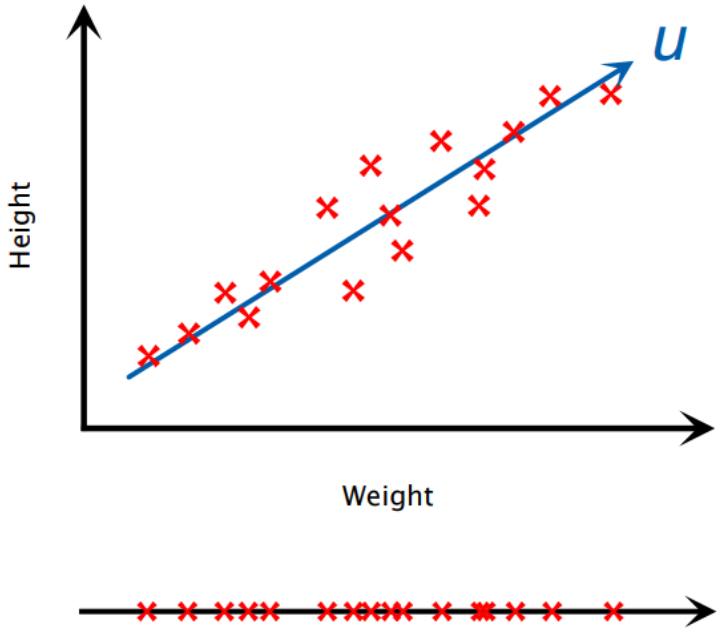
- To understand and analyze high-dimensional dataset in a meaningful way, we would often need to seek a low-dimensional representation of the data to which classical statistical learning tools can be applied.

- A popular approach, named dimension reduction, involves projecting the p variables onto a k-dimensional subspace, where $k < p$.

- This is achieved by computing k different *linear combinations*, or *projections*, of the p variables. Then these k projections are used as predictors to fit various statistical models.

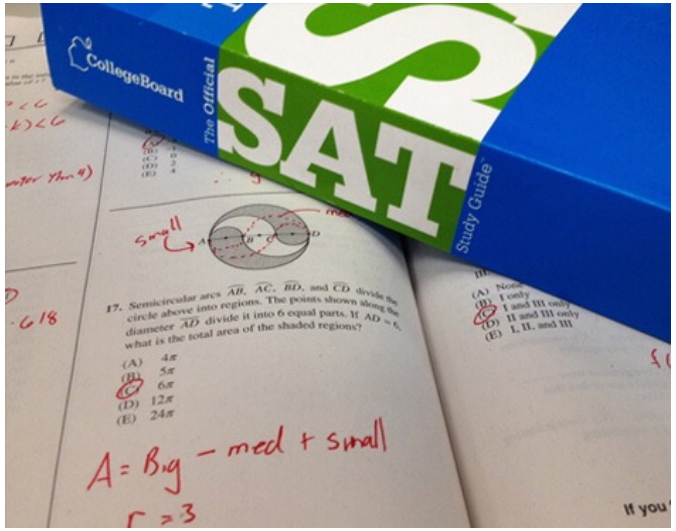
Why Can We Reduce Dimensions?

- Roughly speaking, dimension reduction means we can approximate a p dimensional feature by projecting it onto a lower dimensional space without losing too much information.
- When we have strong correlations between the variables, the data may more or less fall on a line or hyperplane of lower dimensions.



- For instance, imagine a plot of two variables that have a nearly perfect correlation.
- The data points will fall close to a straight line.
- That line could be used as a new (one-dimensional) axis to represent the variation among data points.

Example: SAT Scores



- Here is a toy example: suppose we have verbal, math, and total SAT scores for a group of students.

- We have three variables, but really (at most) two dimensions in the data because

$$\text{total} = \text{verbal} + \text{math},$$

  meaning that the third variable is completely determined by the first two.

- The reason for saying “at most” two dimensions is that if there is a strong correlation between verbal and math, one dimension might just be enough to describe the data.

Principal Component Analysis

- Principal components analysis (PCA) is a widely-used approach for extracting a small set of features from a large set of variables, meanwhile retaining most of the information.

- Principal component analysis is a mathematical procedure that transforms many possibly correlated variables into a smaller number of uncorrelated variables called principal components.

- The first principal component accounts for as much of the variability in the data as possible, and each succeeding principal component accounts for as much of the remaining variability as possible.

Problem Setup

- Suppose we observe a p-dimensional random vector

$$\boldsymbol{x} = (x_1, \dots, x_p)'$$

with population covariance matrix

$$\operatorname{cov}(\boldsymbol{x}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{pmatrix} \in \mathbb{R}^{p \times p}$$

- Consider linear combinations

$$\begin{aligned} Y_1 &= e_{11}x_1 + e_{12}x_2 + \cdots + e_{1p}x_p \\ Y_2 &= e_{21}x_1 + e_{22}x_2 + \cdots + e_{2p}x_p \\ &\vdots \\ Y_p &= e_{p1}x_1 + e_{p2}x_2 + \cdots + e_{pp}x_p \end{aligned}$$

- Each Y_i is a linear combination of $x_1, \dots, x_p$ with coefficients

$$\boldsymbol{e}_i = (e_{i1}, \dots, e_{ip})'$$

Problem Setup (cont.)

- Since Y_i is a function of data, it is also random and has variance

$$\mathrm{var}(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i.$$

- Moreover, Y_i and Y_j have covariance

$$\mathrm{cov}(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{jl} \sigma_{kl} = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_j.$$

Principal component analysis aims to find the linear combination $\boldsymbol{e}_i$ such that $\mathrm{var}(Y_i)$ is maximized.

Scaling the Variables

- As we have discussed in previous lectures, the magnitude of elements in covariance matrix $\boldsymbol{\Sigma}$ can not be directly used as an evidence for the strength of association as it depends on the measure and scale of two variables.

- If you change the unit of length from meter to kilo meter, the value of variable will be 1,000 times smaller. Then the variance of the variable would be tiny, and we may ignore this variable when we look for the direction that maximizes the variance.

- Because it is undesirable for the principal components obtained to depend on an arbitrary choice of scaling, we typically scale each variable to have mean zero and standard deviation one before we perform PCA.

First Principal Component

- The *first principal component* is the linear combination of variables that has the largest variance (among all linear combinations). It accounts for as much variation in the data as possible.

- Specifically we define coefficients $\boldsymbol{e}_1 = (e_{11}, \ldots, e_{1p})'$ for the first principal component as the vector that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} \sigma_{kl} = \boldsymbol{e}_1' \boldsymbol{\Sigma} \boldsymbol{e}_1$$

subject to (as a direction vector)

$$\boldsymbol{e}_1' \boldsymbol{e}_1 = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} = 1.$$

Second Principal Component

- The *second principal component* is the linear combination of variables that accounts for as much of the remaining variation as possible and is uncorrelated with the first component.
- Select $\boldsymbol{e}_2 = (e_{21}, \ldots, e_{2p})'$ for the second principal component that maximizes

$$\text{var}(Y_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{2k} e_{2l} \sigma_{kl} = \boldsymbol{e}_2' \boldsymbol{\Sigma} \boldsymbol{e}_2$$

subject to

$$\boldsymbol{e}_2' \boldsymbol{e}_2 = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{2k} e_{2l} = 1$$

and the additional constraint that these two components are uncorrelated:

$$\text{cov}(Y_1, Y_2) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{2l} \sigma_{kl} = \boldsymbol{e}_1' \boldsymbol{\Sigma} \boldsymbol{e}_2 = 0.$$

i-th Principal Component

- Select $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ip})'$ that maximizes

$$\text{var}(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = \boldsymbol{e}_i' \boldsymbol{\Sigma} \boldsymbol{e}_i$$

subject to

$$\boldsymbol{e}_i' \boldsymbol{e}_i = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} = 1$$

and that this new component is uncorrelated with all the previous ones:

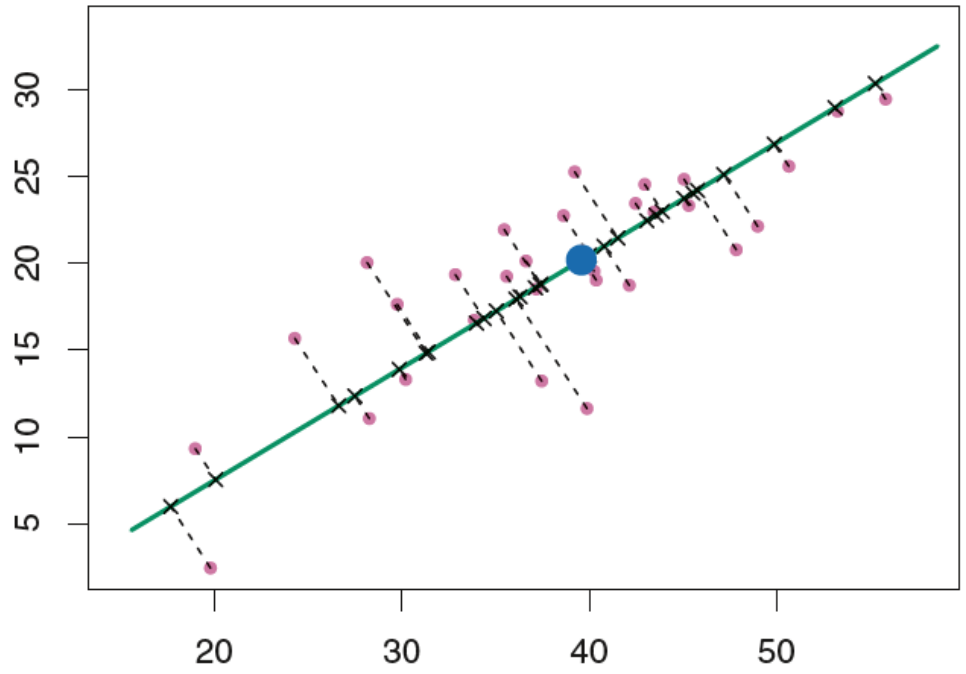
$$\text{cov}(Y_1, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{il} \sigma_{kl} = \boldsymbol{e}_1' \boldsymbol{\Sigma} \boldsymbol{e}_i = 0$$

$$\text{cov}(Y_2, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{2l} \sigma_{kl} = \boldsymbol{e}_2' \boldsymbol{\Sigma} \boldsymbol{e}_i = 0$$

$$\vdots$$

$$\text{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{i-1,k} e_{il} \sigma_{kl} = \boldsymbol{e}_{i-1}' \boldsymbol{\Sigma} \boldsymbol{e}_i = 0$$

Interpretation of First Principal Component

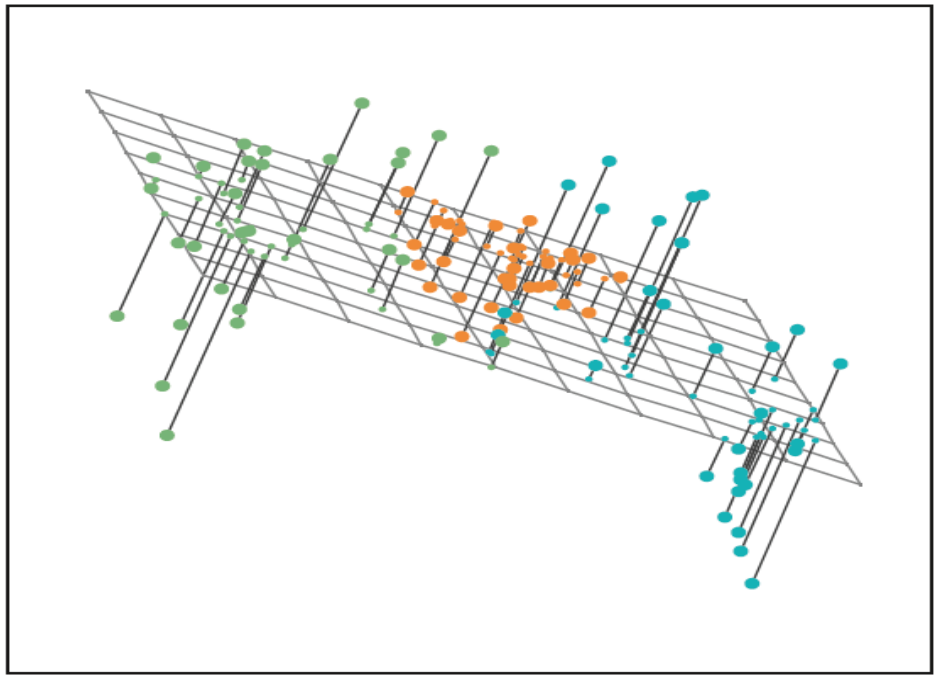


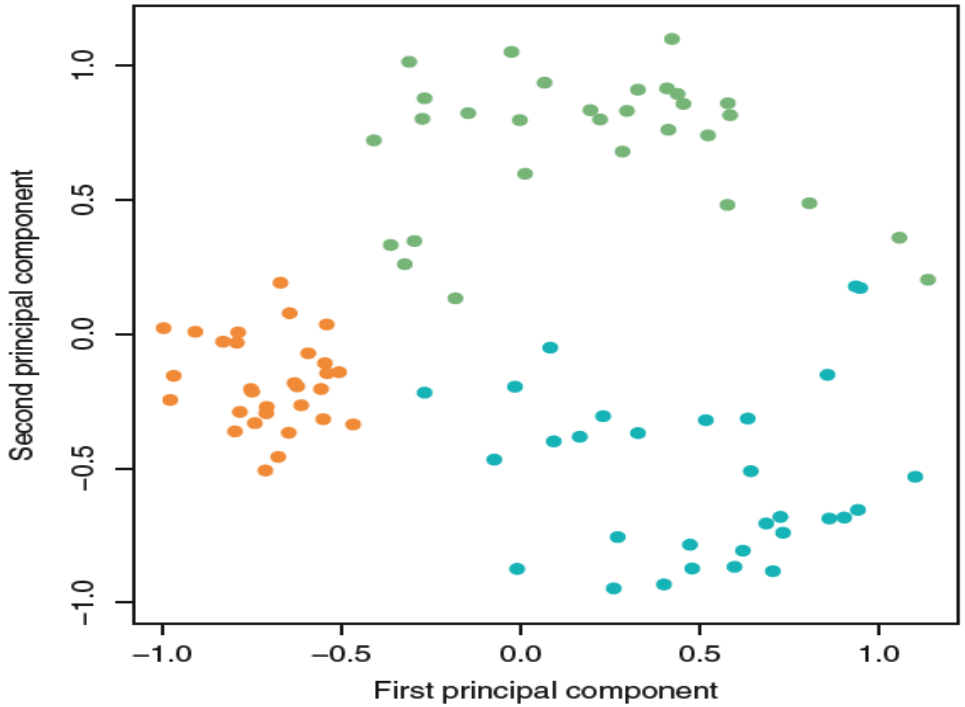
First principal component of a two-dimensional dataset. The blue dot is the sample mean. The green solid line is the first PC. The dashed lines indicate the distance between each observation and the first principal component loading vector.

- The first principal component loading vector has a geometric interpretation: it is the line in $\mathbb{R}^p$ that is *closest* to the *n* observations under the average squared Euclidean distance.

- We seek a single direction along which the univariate projections are as close as possible to all the data points and therefore provide a good summary of the data.

Interpretation of First Two Principal Components





- The first two principal directions span the plane that is closest to the n observations under average squared Euclidean distance.

How to Find Principal Components?

- The principal directions can be found by computing the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$.
- Let $\lambda_1, \dots, \lambda_p$ denote the eigenvalues of $\boldsymbol{\Sigma}$ in descending order, i.e.

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p.$$

- Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ be the corresponding eigenvectors (with unit length):

$$\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \text{ for } i = 1, \dots, p.$$

- It turns out the i-th eigenvector will serve as the i-th principal direction:

1. $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ $(i \neq j)$ due to orthogonality of eigenvectors.
2. $\mathrm{var}(Y_i) = \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_{\mathrm{i}} = \lambda_i \boldsymbol{e}_i'\boldsymbol{e}_i = \lambda_i$. Hence $\mathrm{var}(Y_1) \geq \cdots \geq \mathrm{var}(Y_p)$.
3. $\mathrm{cov}(Y_i, Y_j) = \boldsymbol{e}_i'\Sigma\boldsymbol{e}_{\mathrm{j}} = \lambda_j \boldsymbol{e}_i'\boldsymbol{e}_j = 0, i \neq j.$

Spectral Decomposition Theorem

- The covariance matrix is symmetric and positive semi-definite. It admits the spectral decomposition (e.g. MATH 18):

$$\boldsymbol{\Sigma} = \sum_{j=1}^{p} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j'.$$

- Spectral decomposition states that the linear space spanned by the p features can be fully explained by the p orthogonal directions.

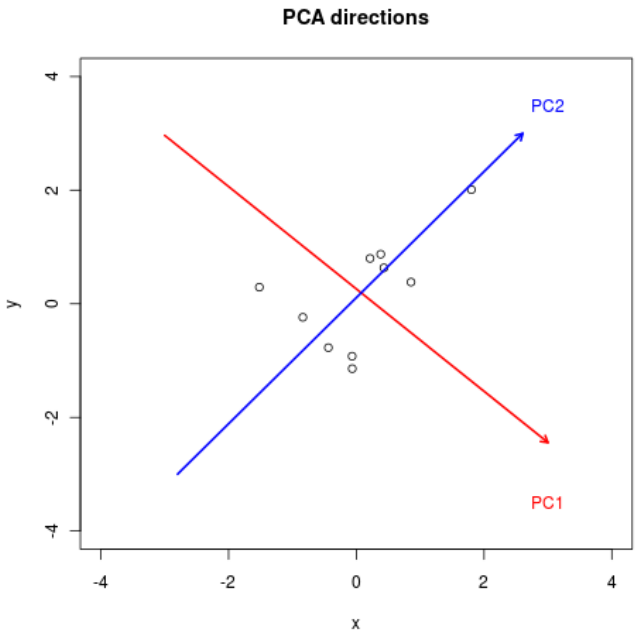
- This decomposition suggests us to approximate $\boldsymbol{\Sigma}$ by its first k eigenvectors:

$$\boldsymbol{\Sigma} \approx \sum_{j=1}^{k} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j' \text{ for some } k < p.$$

- This approximation is useful when $\lambda_{k+1}, \ldots, \lambda_p$ are small. In other words,

$$\boldsymbol{\Sigma} - \sum_{j=1}^{k} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j' = \sum_{j=k+1}^{p} \lambda_j \boldsymbol{e}_j \boldsymbol{e}_j' \text{ is negligible.}$$

Uniqueness of the Principal Components



- Each principal component loading vector is unique, up to a sign flip. This means that two different packages will yield the same principal component loading vectors up to signs.

- The signs may differ because each principal component loading vector specifies a direction in $\mathbb{R}^p$: flipping the sign does not affect the amount of variability contained in each PC.

- In most applications, the sign of principal component loadings will not affect the analysis results. If necessary, one can remove this flexibility by setting the first coordinate of each principal direction to be positive.

How Much Information Do We Loss?

- In week 1 we defined the total variation of $\boldsymbol{X}$ as the trace of the covariance matrix $\boldsymbol{\Sigma}$. This is also the sum of all the eigenvalues:

$$\text{trace}(\boldsymbol{\Sigma}) = \sum_{j=1}^{p} \sigma_j^2 = \sum_{j=1}^{p} \lambda_j \,.$$

- The proportion of variation explained by the i-th principal component is

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

- The proportion of variation explained by the first k principal components is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}.$$

- Naturally, if the proportion of variation explained by the first k principal components is large, then not much information is lost by considering only the first k principal components.

Choose the Number of Principal Components

- To reduce the dimensionality, we only retain the first k principal components.

- Here we need to balance two conflicting purposes:

1. To obtain the simplest possible interpretation, we want k to be as small as possible. If we can explain most of the variation by just looking at the first a few principal components, then it would give us a simple description of the data.

2. To reduce loss of information, we want the proportion of variation explained by the first k principal components to be large (ideally close to 1).

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \approx 1.$$

Estimate Principal Components From a Sample

- Suppose we observe a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of p features.
- The principal components can be estimated via the following steps:

1. Standardize each variable to have mean 0 and standard deviation 1.
2. Calculate the sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\,(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'.$$

3. Calculate eigenvalues and eigenvectors of $\boldsymbol{S}$.
   - Let $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ denote the eigenvalues of $\boldsymbol{S}$ in descending order,
   - Let $\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_p$ be the corresponding eigenvectors.
4. The estimated principal components are

$$\hat{Y}_{i1} = \hat{\boldsymbol{e}}_1'\boldsymbol{x}_i, \;\; \hat{Y}_{i2} = \hat{\boldsymbol{e}}_2'\boldsymbol{x}_i, \ldots, \hat{Y}_{ip} = \hat{\boldsymbol{e}}_p'\boldsymbol{x}_i, \;\; i = 1, \ldots, n.$$

Example: Violent Crime Rates by US State

- This dataset contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. The dataset also includes the percent of the population living in urban areas.

- The dataset contains 50 observations of the following four variables:
  1. Murder: number of murder arrests per 100,000 residents.
  2. Assault: number of assault arrests per 100,000 residents.
  3. Rape: number of rape arrests per 100,000 residents.
  4. UrbanPop: percentage of urban population.

A Peek at the Dataset

- First let's take a quick look at the dataset along with some descriptive statistics.

State	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5

Statistic	Murder	Assault	UrbanPop	Rape
Sample Mean	7.79	170.76	65.54	21.23
Sample SD	18.97	6945.17	209.52	87.73

Principal Component Analysis Results

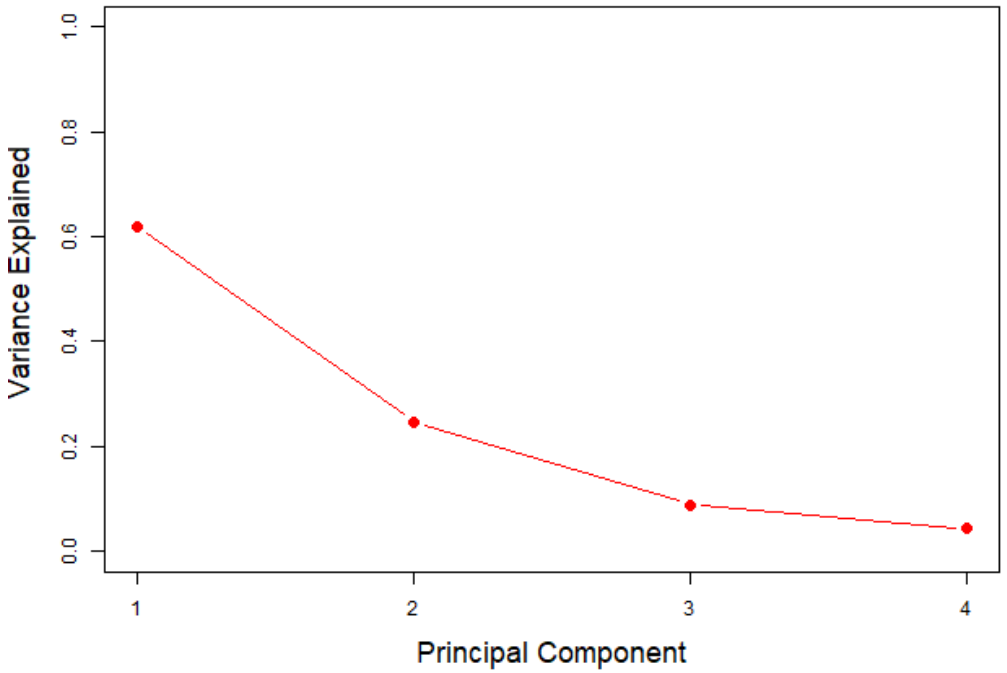
- We first standardize each variable to have mean 0 and standard deviation 1. Then apply PCA to the standardized data.
- All four eigenvalues and their corresponding eigenvectors are listed below.
- We also report the proportion of total variance explained by each PC and the cumulative proportion explained.

Component	Eigenvalue	Proportion	Cumulative
PC 1	2.480	0.620	0.620
PC 2	0.989	0.247	0.867
PC 3	0.357	0.089	0.957
PC 4	0.173	0.043	1.000

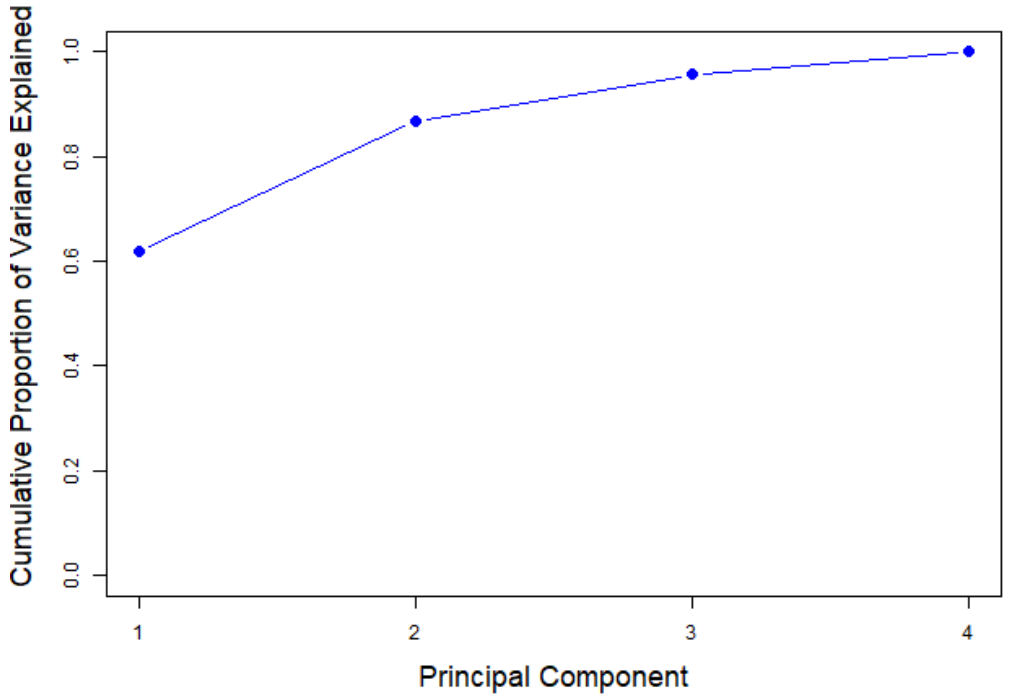
Scree Plot and Cumulative Proportion Plot

- We draw the scree plot to visualize the proportion of total variance explained by each PC. Also, we can visualize the cumulative proportion explained by first k PCs.

Scree Plot

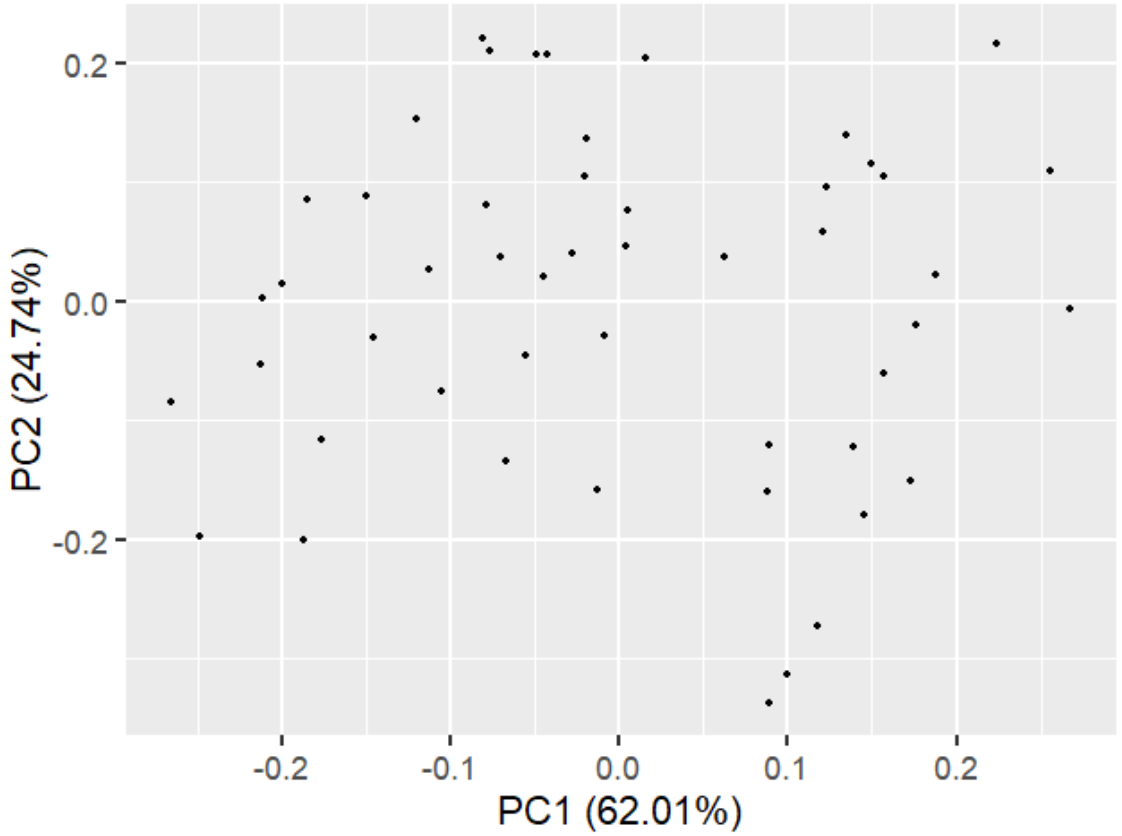


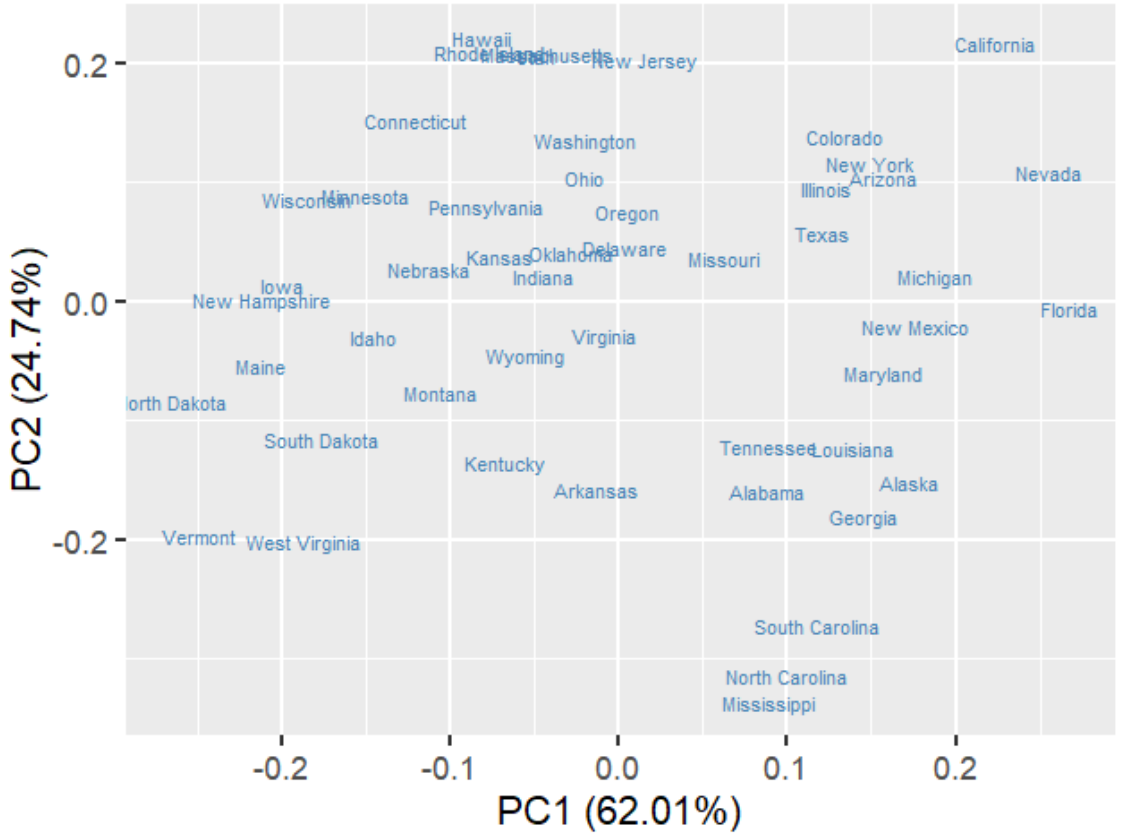
Cumulative Plot



Visualize the Dataset by First Two PCs

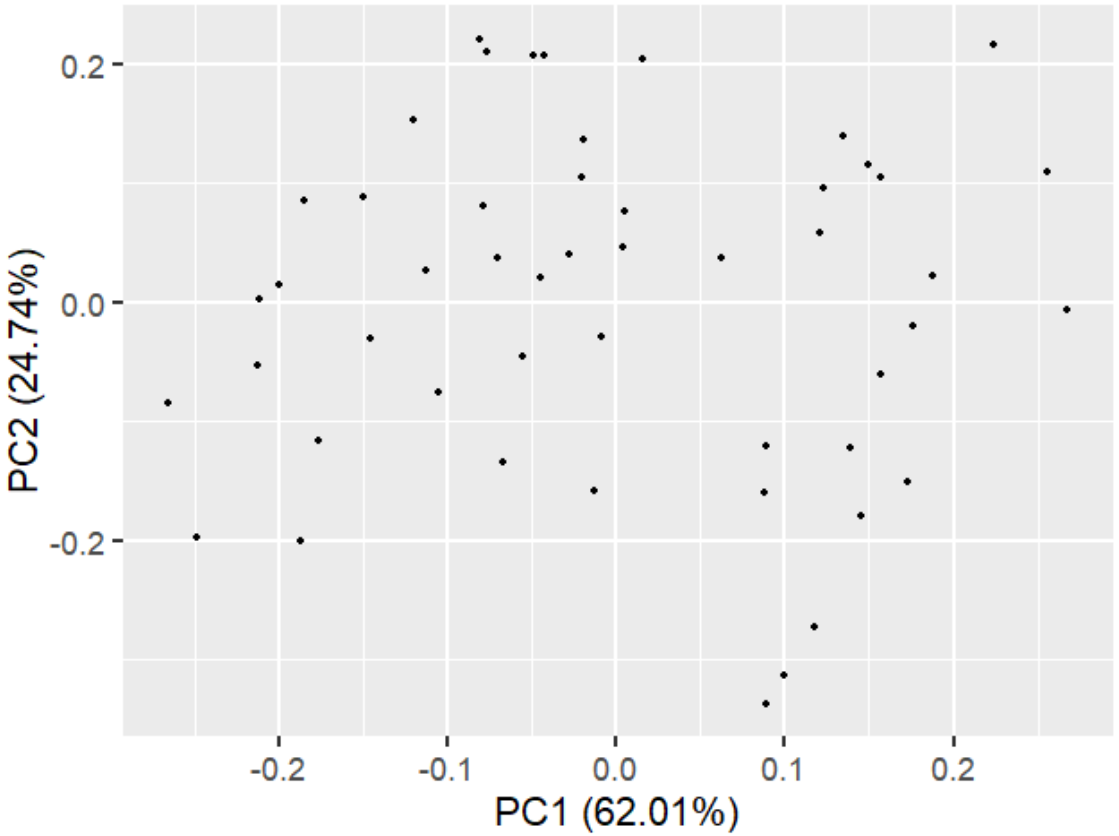
- As a dimension reduction tool, PCA allows us to visualize a multivariate data through a lower dimensional projection.
- Here, we plot the data projected onto the first two PCs.

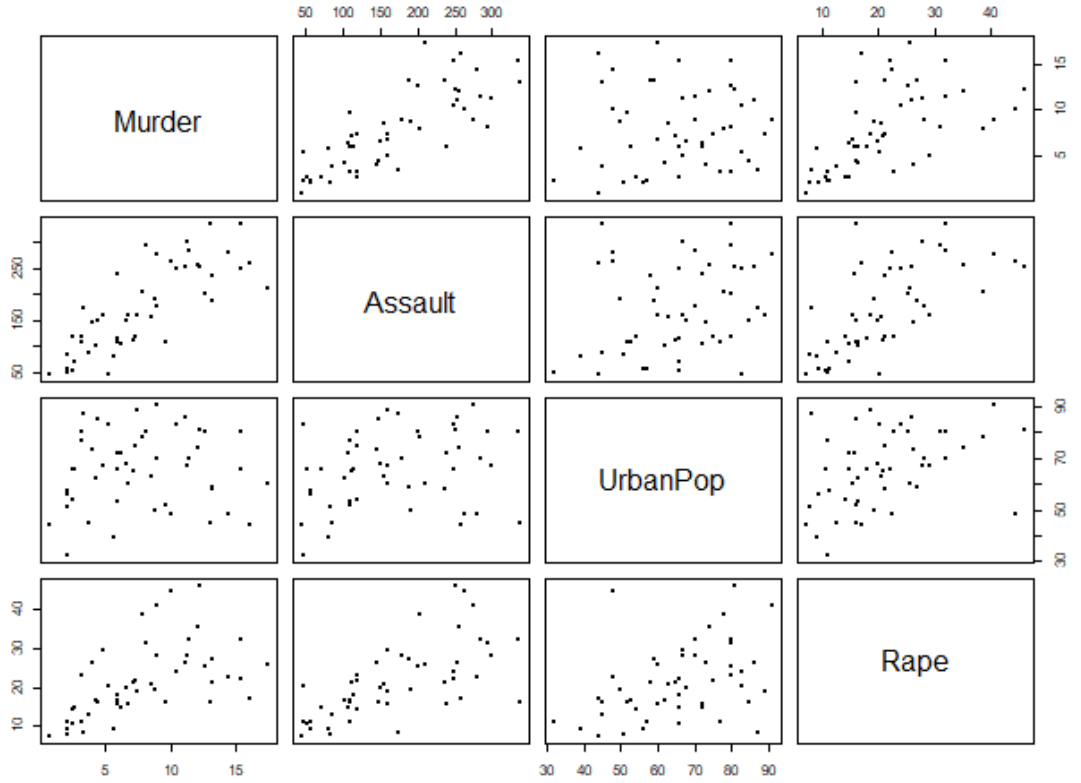




Compare with Pairwise Scatter Plot

- Pairwise scatter plot consists of $p(p-1)/2$ independent scatter plots.
- Pairwise scatter plot views only marginal dependence rather than joint effect.
- Data points are more spread on PC scatter plot.



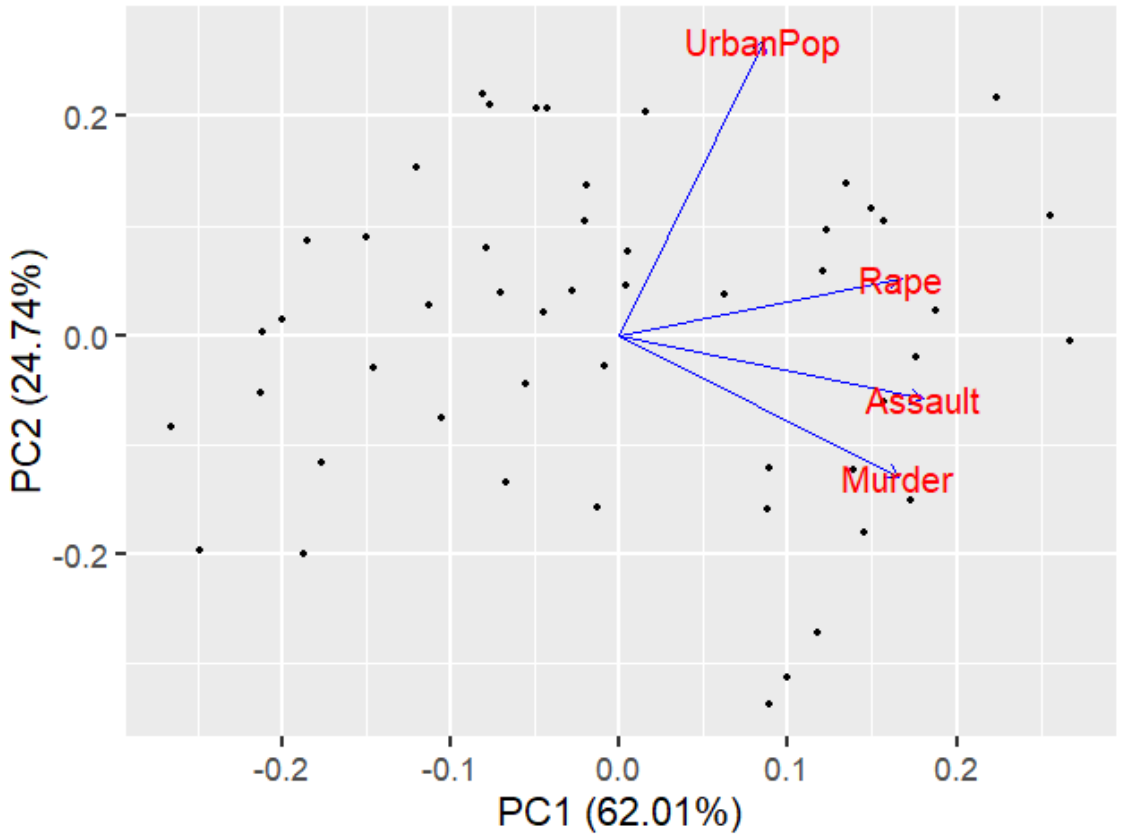


Interpret Principal Components

- The eigenvectors are used as the estimated principal component loading vectors.
- These loading vectors show coefficients of each principal component as a linear combination of variables.
- The calculated principal component coefficients are listed as follows.

	Murder	Assault	UrbanPop	Rape
PC 1	0.536	0.583	0.278	0.543
PC 2	-0.418	-0.188	0.872	0.167
PC 3	0.341	0.268	0.378	-0.818
PC 4	-0.649	0.743	-0.133	-0.089

Visualize Principal Component Loading Vectors

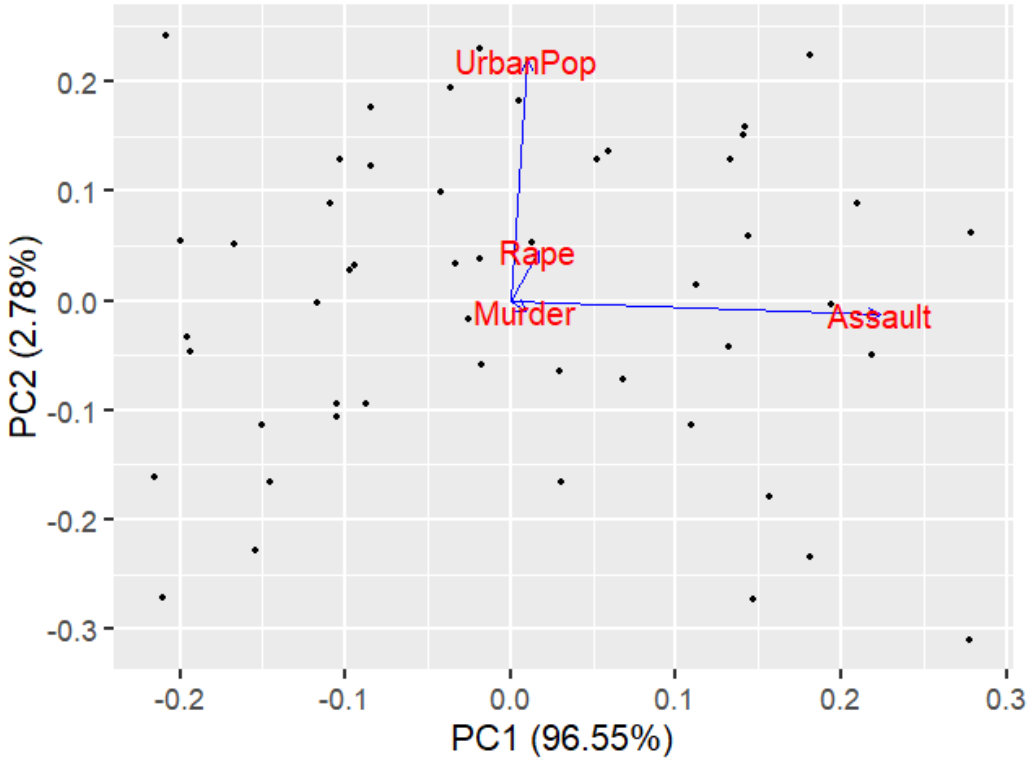


- First loading vector places approximately equal weight on Assault, Murder, and Rape, with much less weight on UrbanPop. Hence this component roughly corresponds to a measure of overall rates of serious crimes.

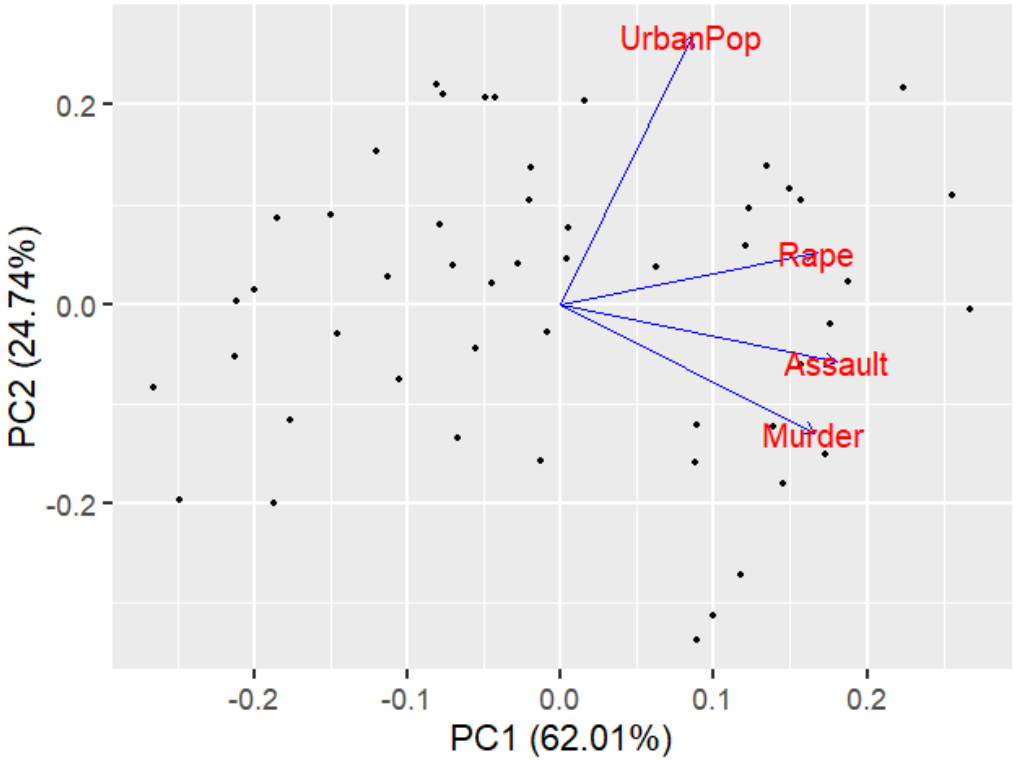
- The second loading vector places most of its weight on UrbanPop and much less weight on the other three features. Hence, this component roughly corresponds to the level of urbanization of the state.

- 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 their correlations.

Raw Data

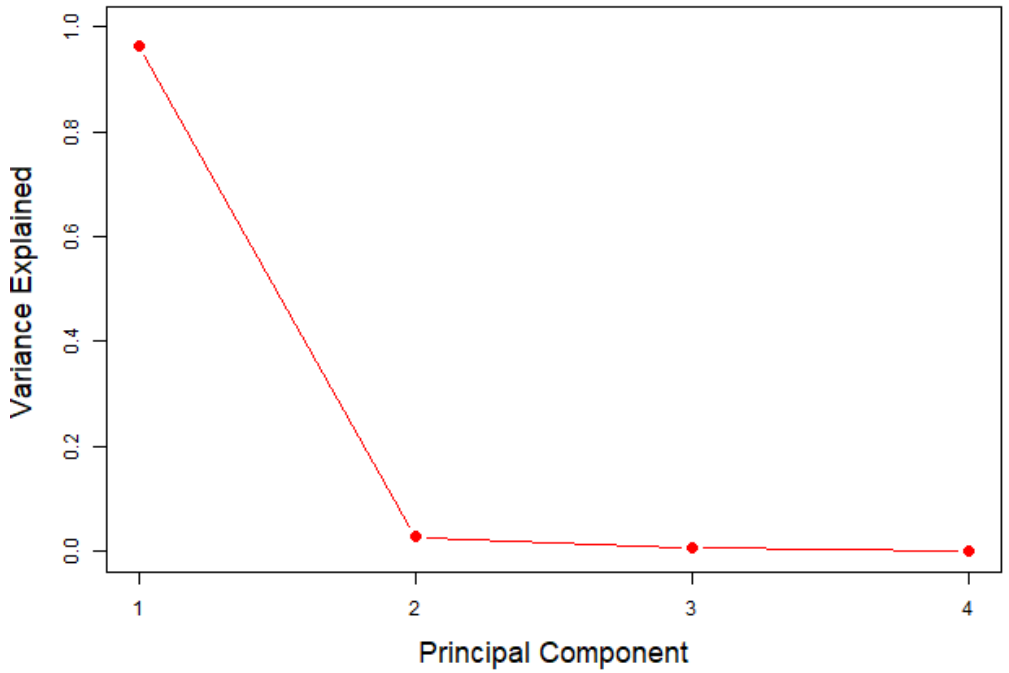


Standardized Data

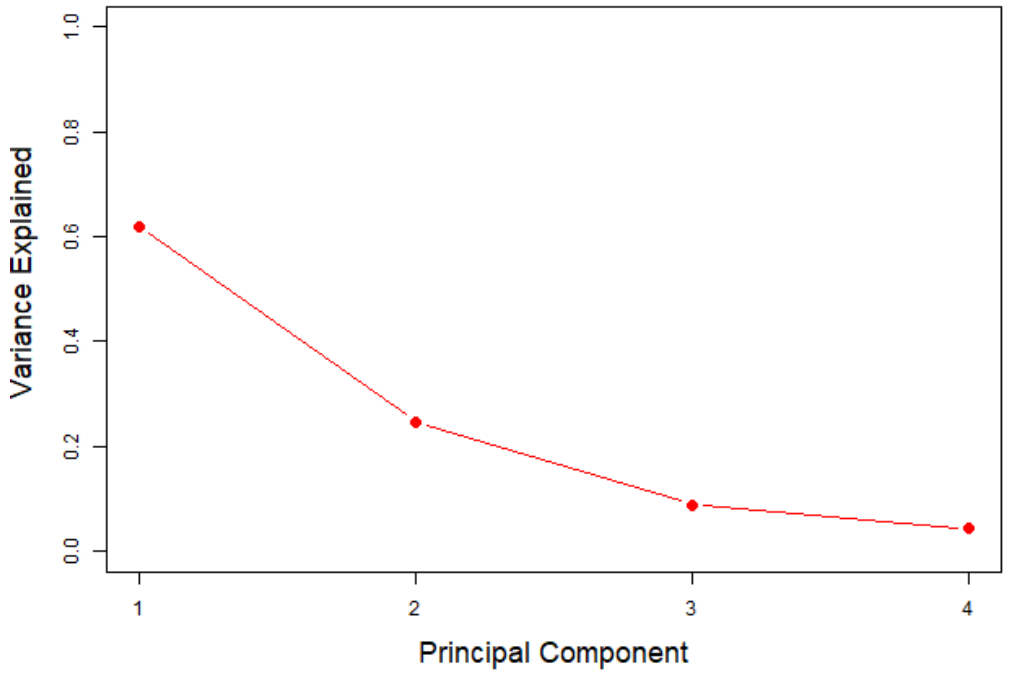


- We also perform PCA on the unscaled data. The first principal component loading vector puts a very large weight on Assault, since this variable has the highest variance.
- The second principal component loading vector places almost all of its weight on UrpanPop. Comparing this to the right-hand plot, we see that scaling does indeed have a substantial effect on the results.

Raw Data



Standardized Data



- When we perform PCA on the unscaled data, the first principal component explains more than 95% of total variation.
- This result is simply a consequence of the scales on which the variables were measured as the total variance is dominated by the variance of Assault.