IF.1105 - IF.2301 Data Science

Lecture 1: Principal Component Analysis (PCA)[1]

Dr. Patricia CONDE-CESPEDES

patricia.conde-cespedes@isep.fr

January 2024

[1] This course is mostly based on... see the references.

Plan

Outline

Introduction to PCA

- **Principal components analysis (PCA)** is a tool used for data visualization or data pre-processing useful for **exploratory** data analysis.

Introduction to PCA

- **Principal components analysis (PCA)** is a tool used for data visualization or data pre-processing useful for **exploratory** data analysis.

Some applications:

- In Biology, for the detection of subgroups of breast cancer patients grouped by their gene expression measurements,
- In marketing for recommender systems
- In social network analysis.

What is PCA?

- Principal Component Analysis *PCA* is a popular approach for **dimensionality reduction**, commonly used for data visualization.

What is PCA?

- Principal Component Analysis *PCA* is a popular approach for **dimensionality reduction**, commonly used for data visualization.
- If the feature-variables are correlated, *PCA* allows to summarize this set with a smaller number of representative variables that collectively explain **most of the variability** in the original set.

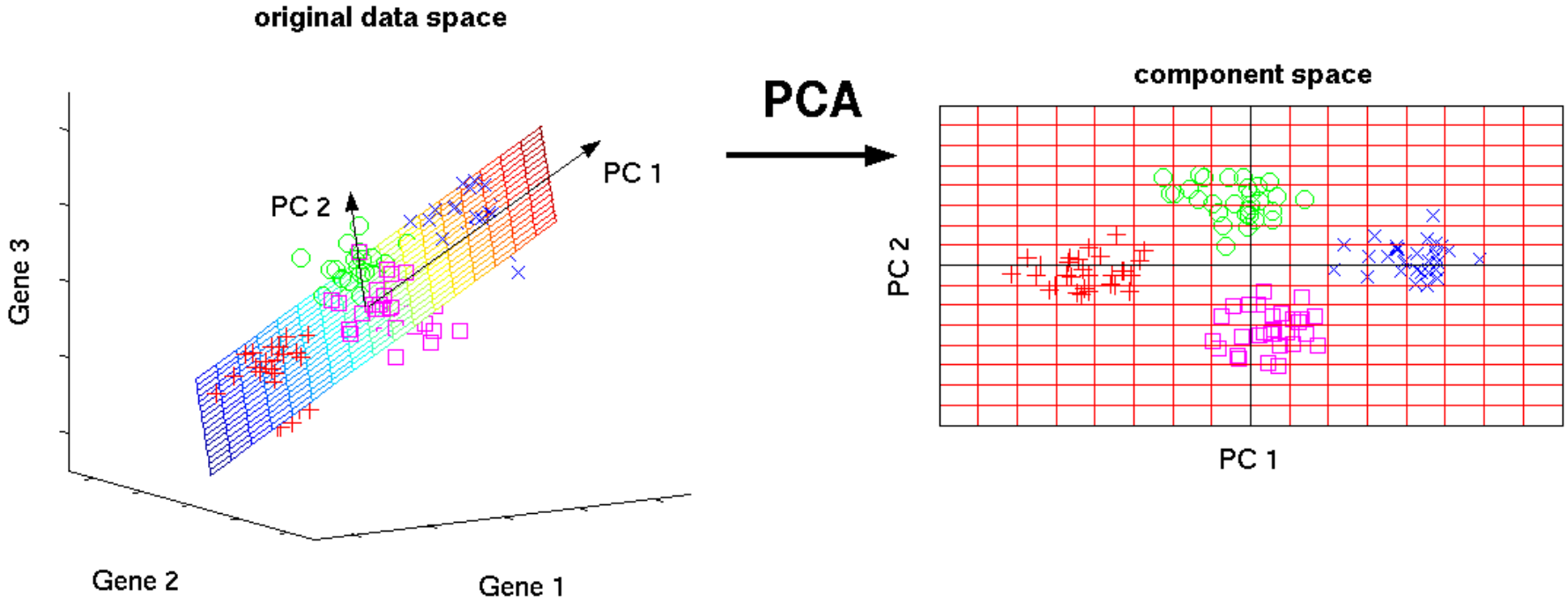
PCA-motivation

How to visualize n observations described by p features?

PCA-motivation

How to visualize n observations described by p features?

Solution $\Rightarrow$ find a low-dimensional representation of the data that captures as much of the information as possible.



This is what *PCA* does.

Source: `https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/`

Outline

The idea behind PCA

Each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting

The idea behind PCA

Each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting .

- *PCA* seeks a small number of dimensions that are as *interesting* as possible.

The idea behind PCA

Each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting .

- *PCA* seeks a small number of dimensions that are as *interesting* as possible.
- We are interested by the amount of variability of observations along each dimension.

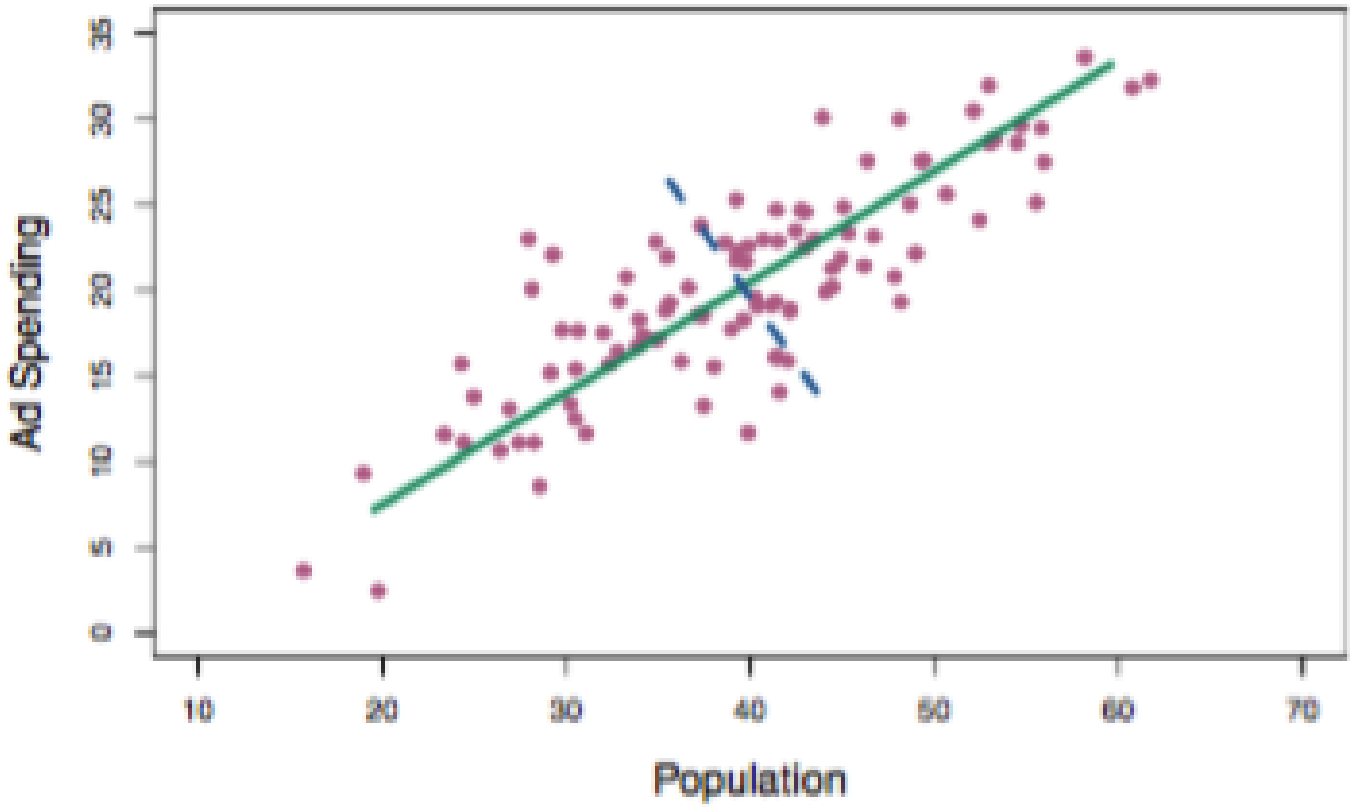
The idea behind PCA

Each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting .

- *PCA* seeks a small number of dimensions that are as *interesting* as possible.
- We are interested by the amount of variability of observations along each dimension.
- Each of the dimensions found by *PCA* is a linear combination of the p original variables.

What are Principal Components?

The first principal component direction represents the direction along which the data vary the most (data is assumed to be centered).



`Advertising data`: population size (pop) in thousands of people, and advertising budget (ad) in thousands of dollars. Green solid line indicates 1st principal component direction; Blue dashed line indicates the 2nd principal component direction.

What is the 1st Principal Component?

- The **first principal component** Z_1 of a set of (centered) features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the original features:

$$Z_1 = \Phi_{11} X_1 + \Phi_{21} X_2 + \ldots + \Phi_{p1} X_p$$

that has the largest variance. Normalized implies $\sum_{j=1}^{p} \Phi_{j1}^2 = 1$.

What is the 1st Principal Component?

- The **first principal component** Z_1 of a set of (centered) features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the original features:

$$Z_1 = \Phi_{11} X_1 + \Phi_{21} X_2 + \ldots + \Phi_{p1} X_p$$

  that has the largest variance. Normalized implies $\sum_{j=1}^{p} \Phi_{j1}^2 = 1$.
- The elements $\Phi_{11}, \ldots, \Phi_{p1}$ are refered to as the **loadings** of the first principal component;

What is the 1st Principal Component?

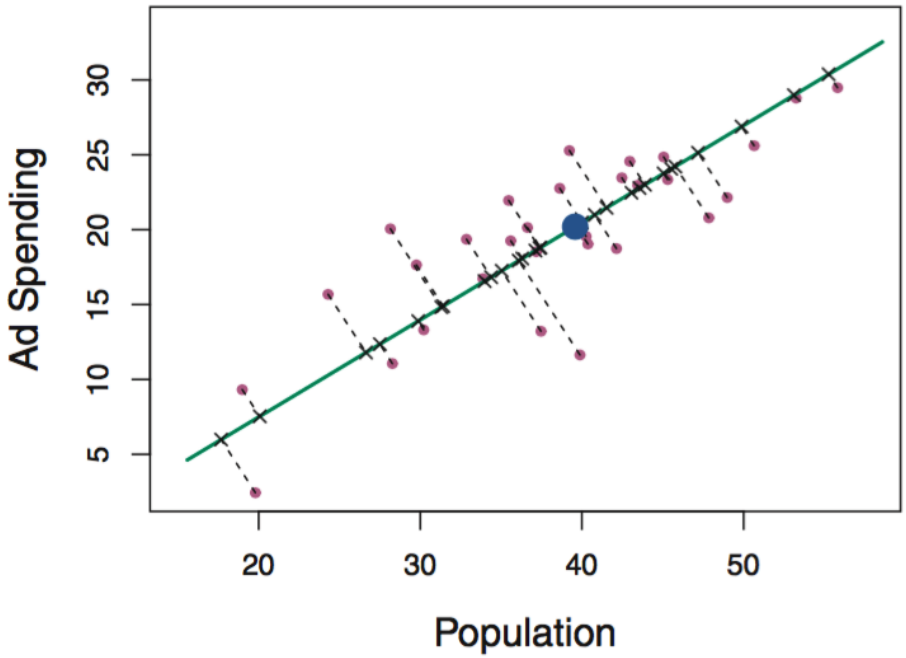
- The **first principal component** Z_1 of a set of (centered) features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the original features:

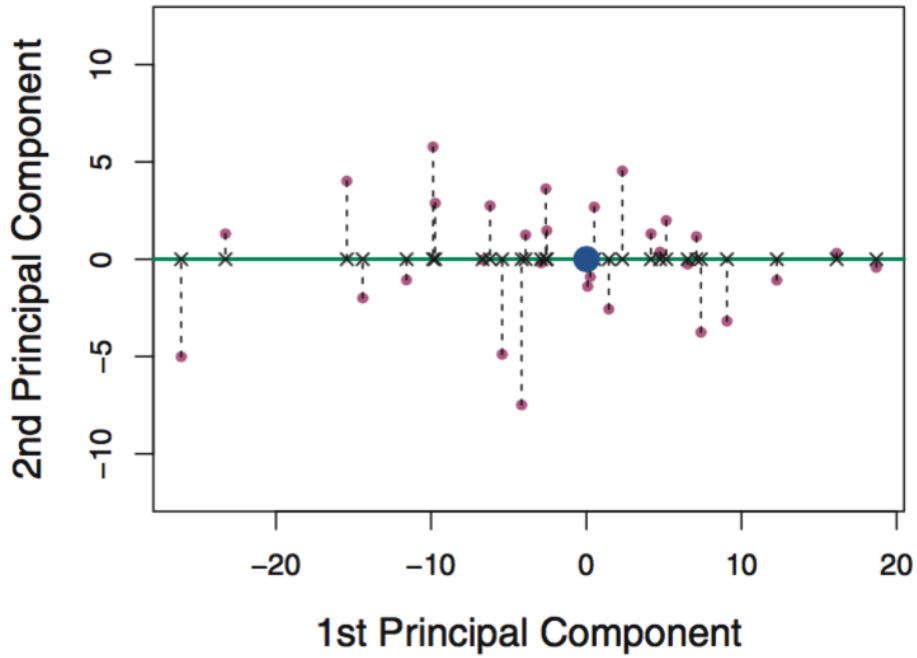
$$Z_1 = \Phi_{11} X_1 + \Phi_{21} X_2 + \ldots + \Phi_{p1} X_p$$

  that has the largest variance. Normalized implies $\sum_{j=1}^{p} \Phi_{j1}^2 = 1$.
- The elements $\Phi_{11}, \ldots, \Phi_{p1}$ are refered to as the **loadings** of the first principal component;
- The vector $\Phi_1 = (\Phi_{11}, \Phi_{21}, \ldots, \Phi_{p1})^T$ is refered to as the first principal component loading vector.

1st Principal Component Interpretation (1/4)

1st Principal component : $Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}})$





Interpretation: The observations projected onto the 1st principal component direction have the largest possible variance. Projecting onto any other line would lead to projected observations with lower variance!

1st Principal Component Interpretation (2/4)

- The *loading* vector, $\Phi_1 = (\Phi_{11}, \Phi_{21})$, defines the direction vector of the first principal component.

1st Principal Component Interpretation (2/4)

- The *loading* vector, $\Phi_1 = (\Phi_{11}, \Phi_{21})$, defines the direction vector of the first principal component.
- **Idea**: Φ_{11} and Φ_{21} are chosen such that:

$$Var(\Phi_{11} \times (pop - \overline{pop}) + \Phi_{21} \times (ad - \overline{ad}))$$

is maximized subject to $\Phi_{11}^2 + \Phi_{21}^2 = 1$.

1st Principal Component Interpretation (2/4)

- The *loading* vector, $\Phi_1 = (\Phi_{11}, \Phi_{21})$, defines the direction vector of the first principal component.
- **Idea**: Φ_{11} and Φ_{21} are chosen such that:

$$Var(\Phi_{11} \times (pop - \overline{pop}) + \Phi_{21} \times (ad - \overline{ad}))$$

  is maximized subject to $\Phi_{11}^2 + \Phi_{21}^2 = 1$.

- Z_1 is a vector of size n. Given the observation i, the value $z_{i1} = 0.839 \times (\text{pop}_i - \overline{\text{pop}}) + 0.544 \times (\text{ad}_i - \overline{\text{ad}})$ is known as the **first principal component score of observation** i.

1st Principal Component Interpretation (3/4)

More interpretations:

- The first principal component vector defines the line that is *as close as possible* to the data. The sum of the squared perpendicular distances between each point and the line is minimal.

1st Principal Component Interpretation (3/4)

More interpretations:

- The first principal component vector defines the line that is *as close as possible* to the data. The sum of the squared perpendicular distances between each point and the line is minimal.
- The first principal component score for the ith observation represents the distance of the ith projection from zero. In other words, the coordinates of i in the new axis.

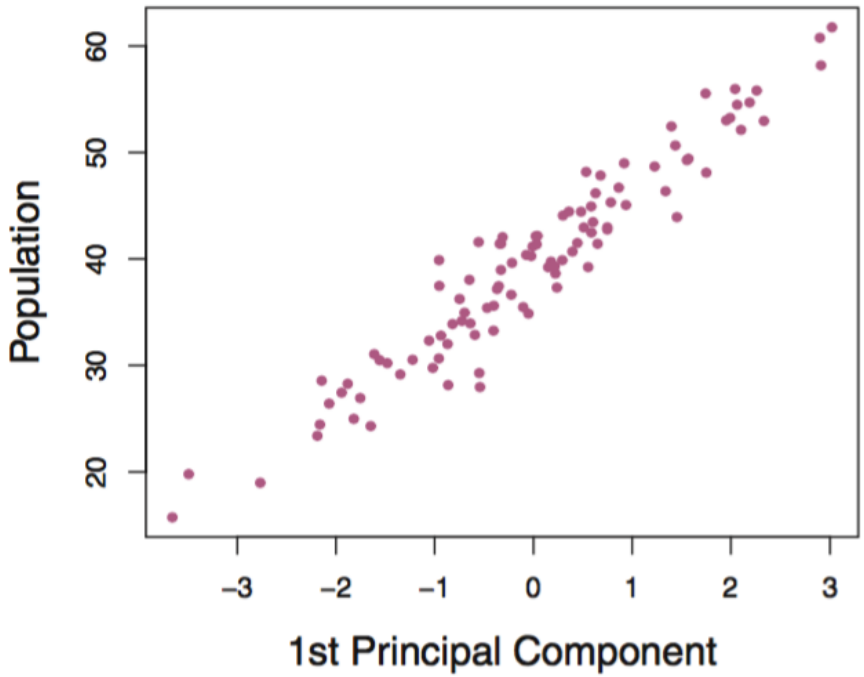
1st Principal Component Interpretation (3/4)

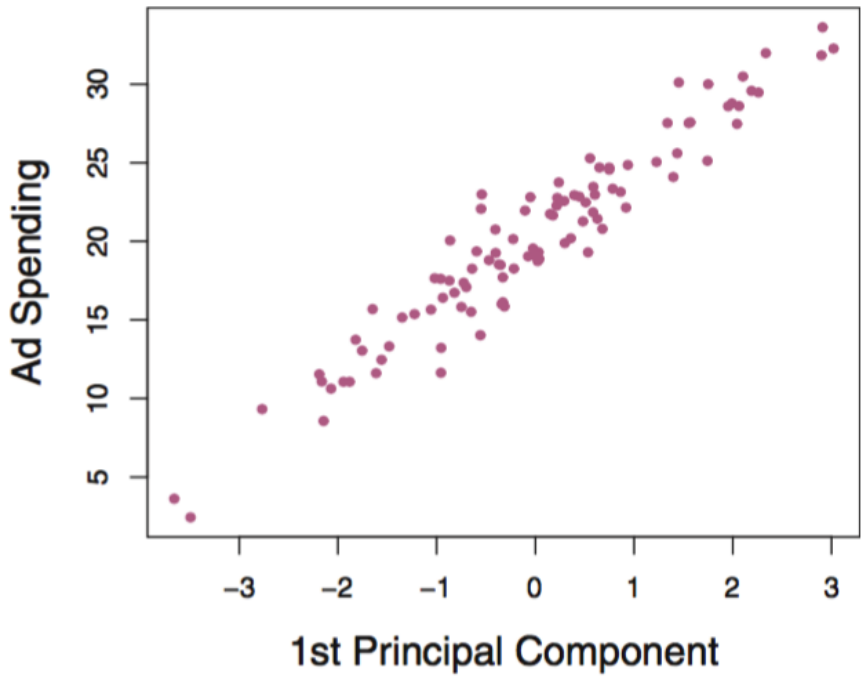
More interpretations:

- The first principal component vector defines the line that is *as close as possible* to the data. The sum of the squared perpendicular distances between each point and the line is minimal.
- The first principal component score for the ith observation represents the distance of the ith projection from zero. In other words, the coordinates of i in the new axis.
- The values of the principal component Z_1 can be interpreted as single-number summaries of both variables pop and ad.

1st Principal Component Interpretation (4/4)

How can a single number represent both variables pop and ad?





The first Principal component versus pop and Ad.

pop and ad have approximately a linear relationship, so a single-number can approximately summarize both.

The 1st principal component appears to capture most of the information contained in both predictors.

How to compute the 1st Principal Component?

- Suppose we have a $n \times p$ data set $\mathbf{X}$. All the variables in $\mathbf{X}$ have been centered to have **mean zero**.

How to compute the 1st Principal Component?

- Suppose we have a $n \times p$ data set $\mathbf{X}$. All the variables in $\mathbf{X}$ have been centered to have **mean zero**.
- We then look for the linear combination of the form:

$$z_{i1} = \Phi_{11} x_{i1} + \Phi_{21} x_{i2} + \ldots + \Phi_{p1} x_{ip} \quad \forall i = 1, \ldots, n$$

that has largest sample variance, subject to $\sum_{j=1}^{p} \Phi_{j1}^2 = 1$.

How to compute the 1st Principal Component?

- Suppose we have a $n \times p$ data set $\mathbf{X}$. All the variables in $\mathbf{X}$ have been centered to have **mean zero**.
- We then look for the linear combination of the form:

$$z_{i1} = \Phi_{11} x_{i1} + \Phi_{21} x_{i2} + \ldots + \Phi_{p1} x_{ip} \quad \forall i = 1, \ldots, n$$

that has largest sample variance, subject to $\sum_{j=1}^{p} \Phi_{j1}^2 = 1$.

- Since each X_j has mean zero, then so does Z_1 (for any values ϕ_{j1}). Hence the sample variance of the z_{i1} can be written as $\frac{1}{n} \sum_{i=1}^{n} z_{i1}^2$.

1st Principal Component optimization problem

Then, formaly the first principal component loading vector is the solution of the following optimization problem:

$$\underset{\Phi_{11},\ldots,\Phi_{p1}}{\text{maximize}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1} x_{ij} \right)^2 \right) \text{ subject to } \sum_{j=1}^{p} \Phi_{j1}^2 = 1.$$

- It can be shown that this problem can be solved via an eigen value decomposition of the variance matrix $\mathbf{X}^T\mathbf{X}$ (centered).

1st Principal Component optimization problem

Then, formaly the first principal component loading vector is the solution of the following optimization problem:

$$\underset{\Phi_{11},\ldots,\Phi_{p1}}{\text{maximize}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1} x_{ij} \right)^2 \right) \text{ subject to } \sum_{j=1}^{p} \Phi_{j1}^2 = 1.$$

- It can be shown that this problem can be solved via an eigen value decomposition of the variance matrix $\mathbf{X}^T\mathbf{X}$ (centered).
- The first principal component loading vector $\Phi_1 = (\Phi_{11}, \Phi_{21}, \ldots, \Phi_{p1})$ defines a direction in feature space along which the data vary the most.

1st Principal Component optimization problem

Then, formaly the first principal component loading vector is the solution of the following optimization problem:

$$\underset{\Phi_{11},\ldots,\Phi_{p1}}{\text{maximize}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1} x_{ij} \right)^2 \right) \text{ subject to } \sum_{j=1}^{p} \Phi_{j1}^2 = 1.$$

- It can be shown that this problem can be solved via an eigen value decomposition of the variance matrix $\mathbf{X}^T\mathbf{X}$ (centered).
- The first principal component loading vector $\Phi_1 = (\Phi_{11}, \Phi_{21}, \ldots, \Phi_{p1})$ defines a direction in feature space along which the data vary the most.
- Φ_1 is the eigen vector associated to the largest eigen value of $\mathbf{X}^T\mathbf{X}$.

1st Principal Component optimization problem

Then, formaly the first principal component loading vector is the solution of the following optimization problem:

$$\underset{\Phi_{11},\ldots,\Phi_{p1}}{\text{maximize}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \Phi_{j1} x_{ij} \right)^2 \right) \text{ subject to } \sum_{j=1}^{p} \Phi_{j1}^2 = 1.$$

- It can be shown that this problem can be solved via an eigen value decomposition of the variance matrix $\mathbf{X}^T\mathbf{X}$ (centered).
- The first principal component loading vector $\Phi_1 = (\Phi_{11}, \Phi_{21}, \ldots, \Phi_{p1})$ defines a direction in feature space along which the data vary the most.
- Φ_1 is the eigen vector associated to the largest eigen value of $\mathbf{X}^T\mathbf{X}$.
- the principal component scores $z_{11}, \ldots, z_{n1}$ are the projections of the n data points $x_1, \ldots, x_n$ onto this direction.

What about the calculation of further Principal Components?

- The second principal component Z_2 corresponds to the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are **uncorrelated** with Z_1.

What about the calculation of further Principal Components?

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

$$z_{i2} = \Phi_{12} x_{i1} + \Phi_{22} x_{i2} + \ldots + \Phi_{p2} x_{ip}, \text{for observation } i$$

The vector $\Phi_2 = (\Phi_{12}, \Phi_{22}, \ldots, \Phi_{p2})$ is refered to as the second principal component loading vector.

What about the calculation of further Principal Components?

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

$$z_{i2} = \Phi_{12} x_{i1} + \Phi_{22} x_{i2} + \ldots + \Phi_{p2} x_{ip}, \text{for observation } i$$

  The vector $\Phi_2 = (\Phi_{12}, \Phi_{22}, \ldots, \Phi_{p2})$ is refered to as the second principal component loading vector.
- It turns out that: Z_2 uncorrelated with $Z_1 \Leftrightarrow \Phi_2$ **orthogonal** to Φ_1.

What about the calculation of further Principal Components?

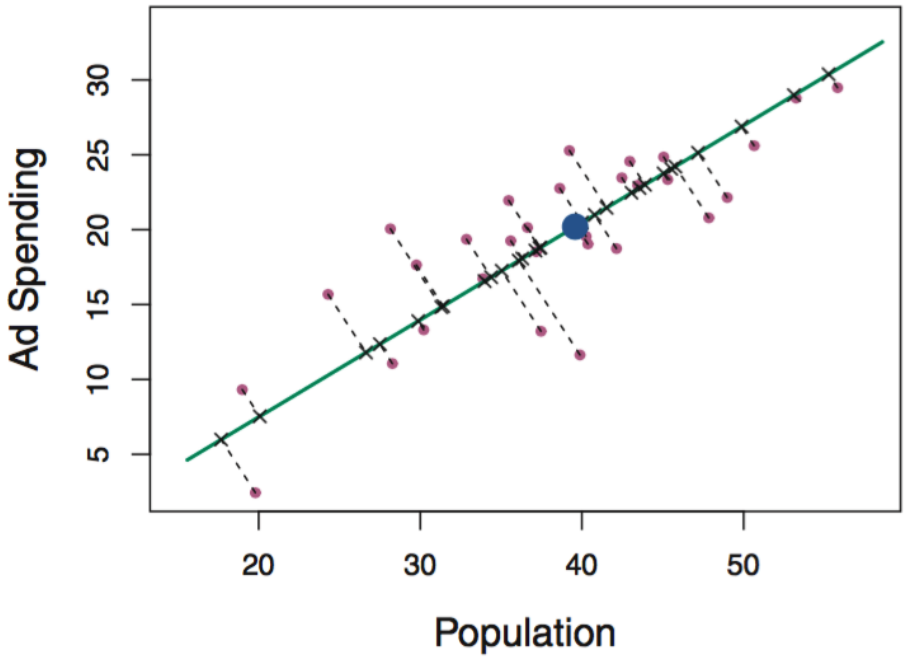
- The second principal component Z_2 corresponds to the linear combination of $X_1, \ldots, X_p$ that has maximal variance among all linear combinations that are **uncorrelated** with Z_1.
- The second principal component scores $z_{12}, z_{22}, \ldots, z_{n2}$ take the form:

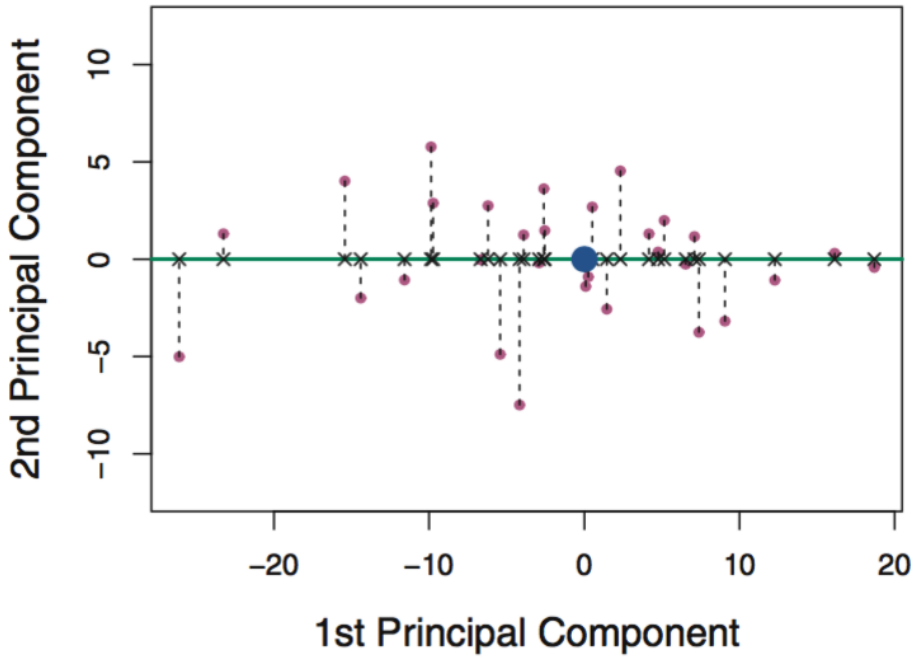
$$z_{i2} = \Phi_{12} x_{i1} + \Phi_{22} x_{i2} + \ldots + \Phi_{p2} x_{ip}, \text{for observation } i$$

  The vector $\Phi_2 = (\Phi_{12}, \Phi_{22}, \ldots, \Phi_{p2})$ is refered to as the second principal component loading vector.
- It turns out that: Z_2 uncorrelated with $Z_1 \Leftrightarrow \Phi_2$ **orthogonal** to Φ_1.
- Φ_2 is the eigen vector of $\mathbf{X}^T\mathbf{X}$ associated to the second largest eigen vector.
- Analogously, one can obtain the other principal component directions $\Phi_3, \Phi_4, \ldots \Phi_p$.

Example: 1st and 2nd Principal components, `Advertising` data

2nd Principal component : $Z_2 = 0.544 \times (\text{pop} - \overline{\text{pop}}) - 0.839 \times (\text{ad} - \overline{\text{ad}})$

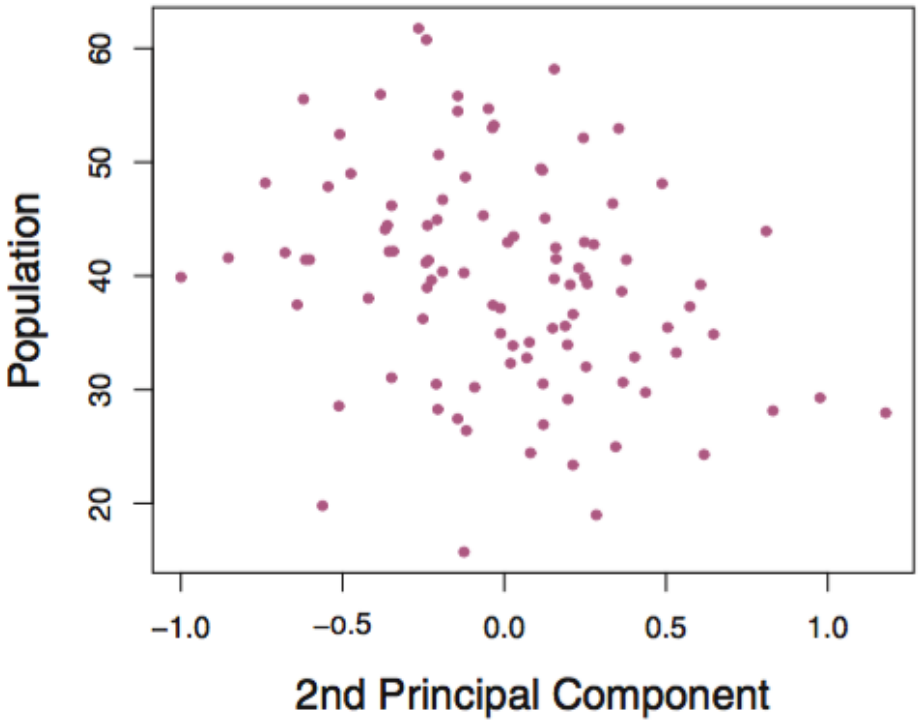


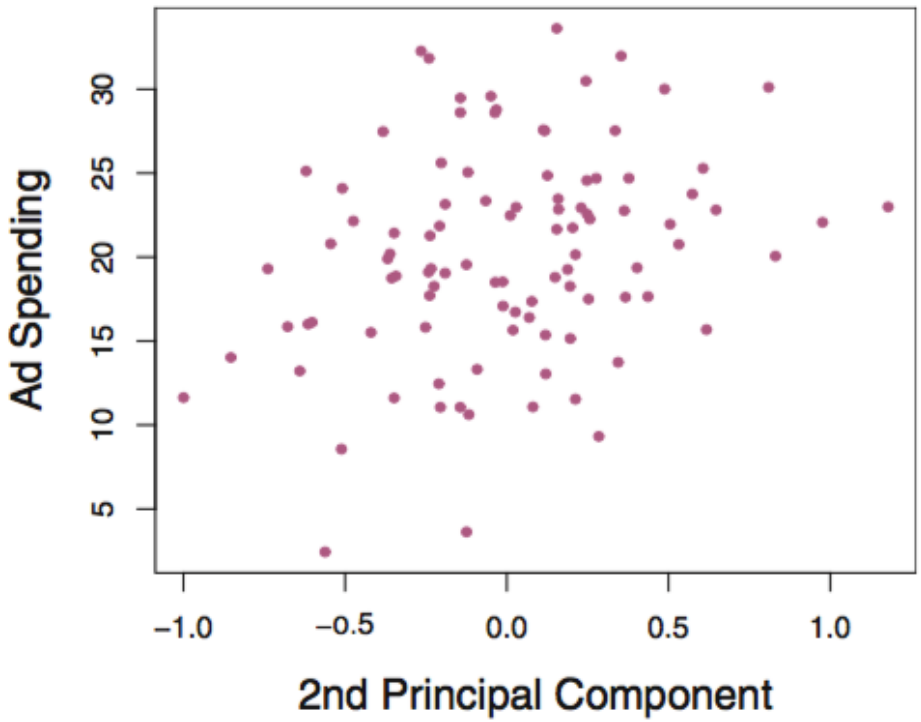


The 2nd Principal Component captures much less information than the 1st Principal Component!

Relationship between the 2nd Principal Component and the features

There is little relationship between the second principal component and the two variables!





Then, using only one dimension representation would be reasonable!

This is why PCA is a visualization method for dimensionality reduction!

ACP application to `USArrests` data set (1/4)

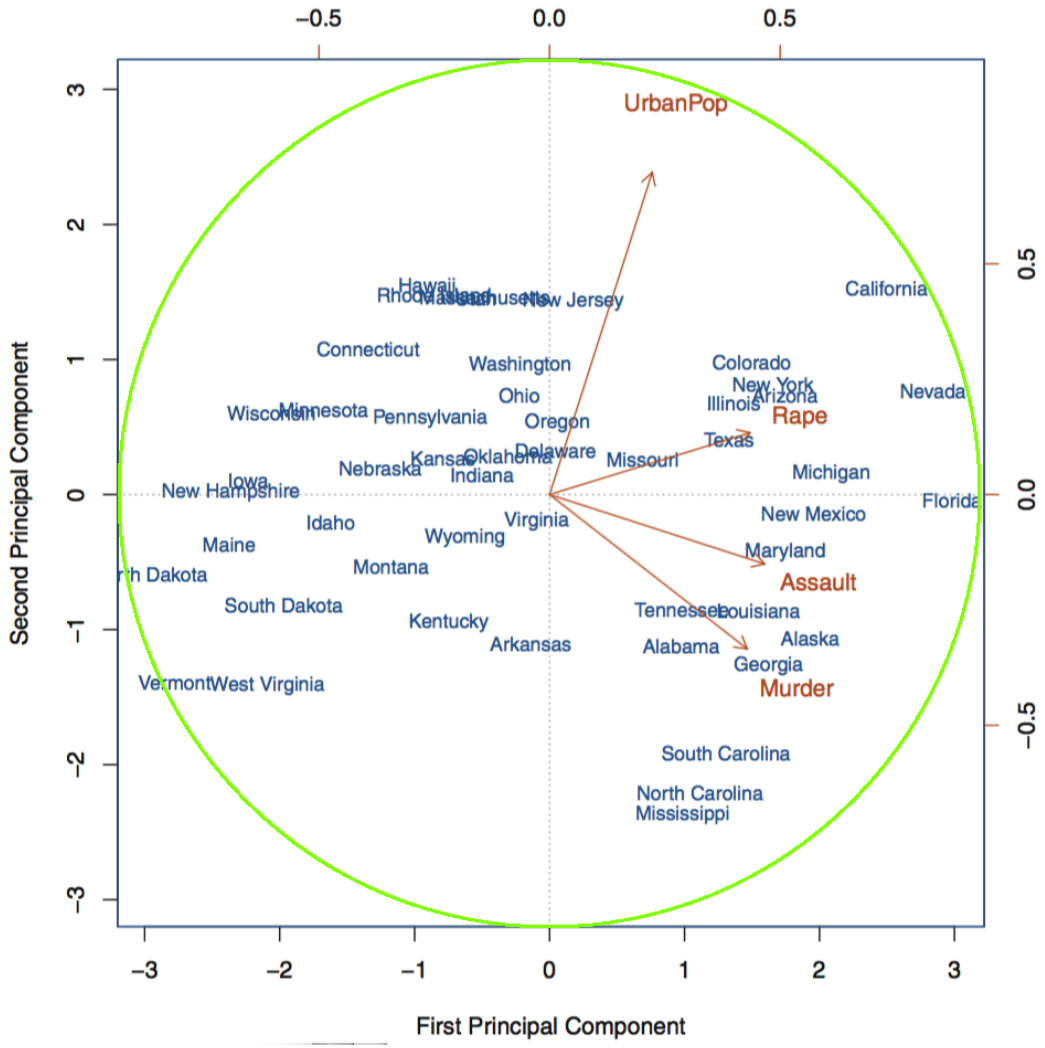
- For each of the 50 states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three types of crimes: `Assault`, `Murder`, and `Rape`. There si also `UrbanPop` (the percent of the population in each state living in urban areas).
- The data set contains $n = 50$ observations described by $p = 4$ variables.

ACP application to `USArrests` data set (1/4)

- For each of the 50 states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three types of crimes: `Assault`, `Murder`, and `Rape`. There si also `UrbanPop` (the percent of the population in each state living in urban areas).
- The data set contains $n = 50$ observations described by $p = 4$ variables.
- PCA was performed after standardizing each variable, that is to have mean zero and standard deviation one.

ACP application to `USArrests` data set (2/4)

The biplot of the first two principal components.



Loadings:

	PC1	PC2
Murder	0.53	-0.42
Assault	0.58	-0.19
UrbanPop	0.28	0.87
Rape	0.54	0.17

ACP application to `USArrests` data set (3/4)

Details on the previous results:

- The figure is called **biplot** because it displays both the principal component scores and the principal component loadings.
- The blue state names represent the scores for the first two principal components (with axes on the bottom and left).

ACP application to `USArrests` data set (3/4)

Details on the previous results:

- The figure is called **biplot** because it displays both the principal component scores and the principal component loadings.
- The blue state names represent the scores for the first two principal components (with axes on the bottom and left).
- The red 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)].

ACP application to `USArrests` data set (3/4)

Details on the previous results:

- The figure is called **biplot** because it displays both the principal component scores and the principal component loadings.
- The blue state names represent the scores for the first two principal components (with axes on the bottom and left).
- The red 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)].
- The green circle is called correlation circle of radius 1.

ACP application to `USArrests` data set (4/4)

Remarks on the biplot:

- The first loading vector places approximately equal weight on `Assault`, `Murder`, and `Rape`, with much less weight on `UrbanPop`. Hence this component roughly measures of overall rates of serious crimes.

ACP application to `USArrests` data set (4/4)

Remarks on the biplot:

- The first loading vector places approximately equal weight on `Assault`, `Murder`, and `Rape`, with much less weight on `UrbanPop`. Hence this component roughly measures 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.

ACP application to `USArrests` data set (4/4)

Remarks on the biplot:

- The first loading vector places approximately equal weight on `Assault`, `Murder`, and `Rape`, with much less weight on `UrbanPop`. Hence this component roughly measures 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 `UrbanPop` is far from them. This indicates that the crime-related variables are correlated with each other.

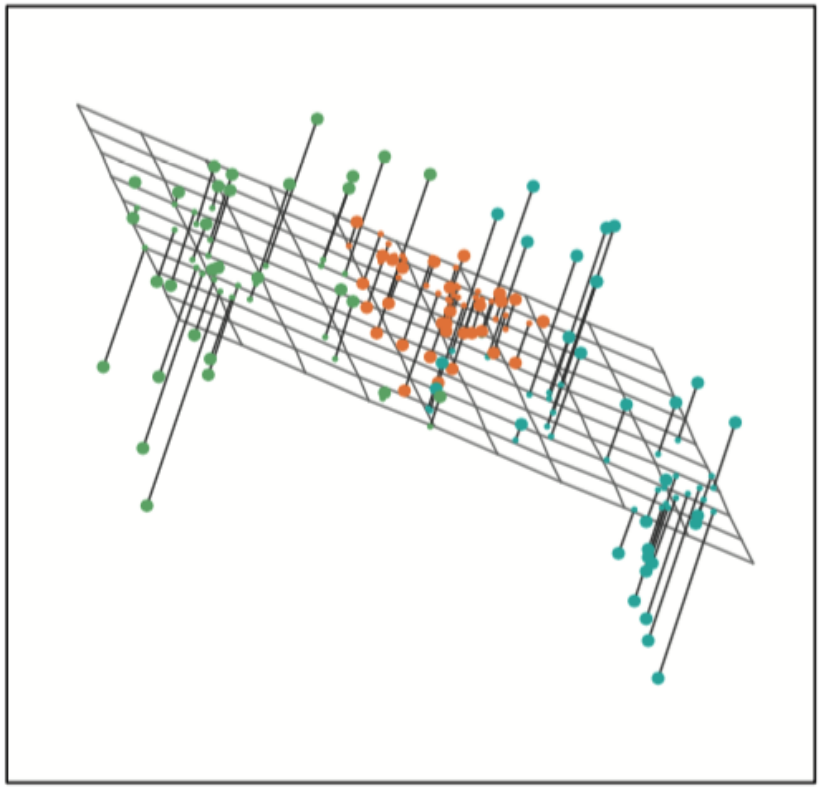
ACP application to `USArrests` data set (4/4)

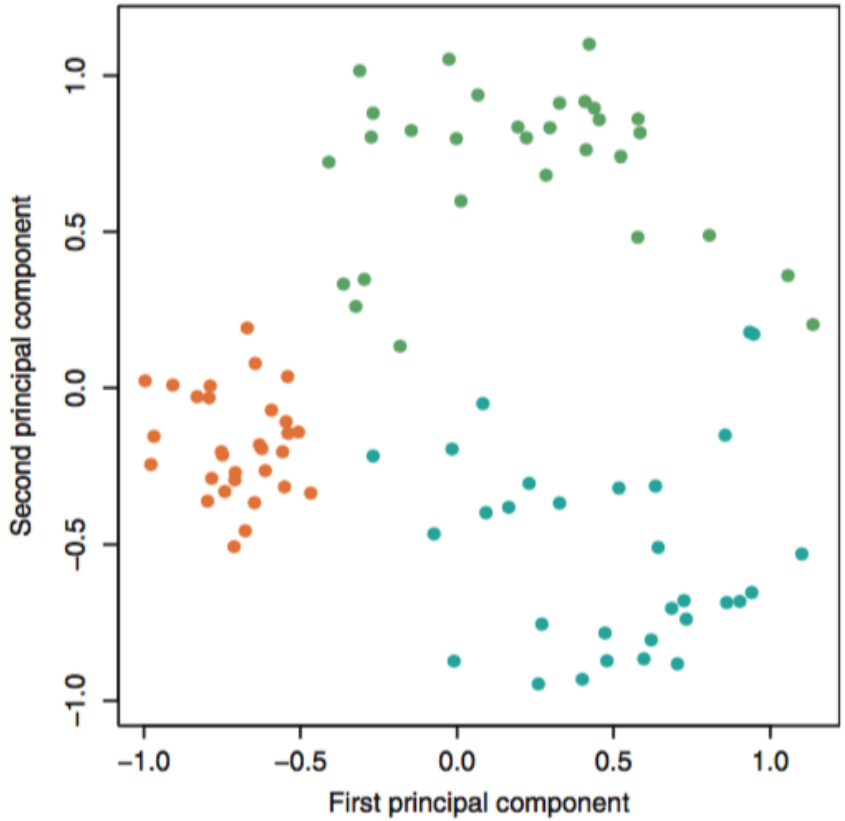
Remarks on the biplot:

- The first loading vector places approximately equal weight on `Assault`, `Murder`, and `Rape`, with much less weight on `UrbanPop`. Hence this component roughly measures 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 `UrbanPop` is far from them. This indicates that the crime-related variables are correlated with each other.
- States with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states 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.

Another interpretation of Principal Components

The first two principal component loading vectors in a simulated three-dimensional data set:





Principal components provide low-dimensional linear surfaces that are as close as possible to the observations.

Another use of PCA: data approximation

- The first principal component is the line in p-dimensional space that is closest to the n observations (using Euclidean distance).

Another use of PCA: data approximation

- The first principal component is the line in p-dimensional space that is closest to the n observations (using Euclidean distance).
- The notion can be extended. The first 2 principal components span the plane that is closest to the n observations...
- The first 3 principal components span the 3D dimensional hyperplane that is closest to the n observations, and...
- so forth.

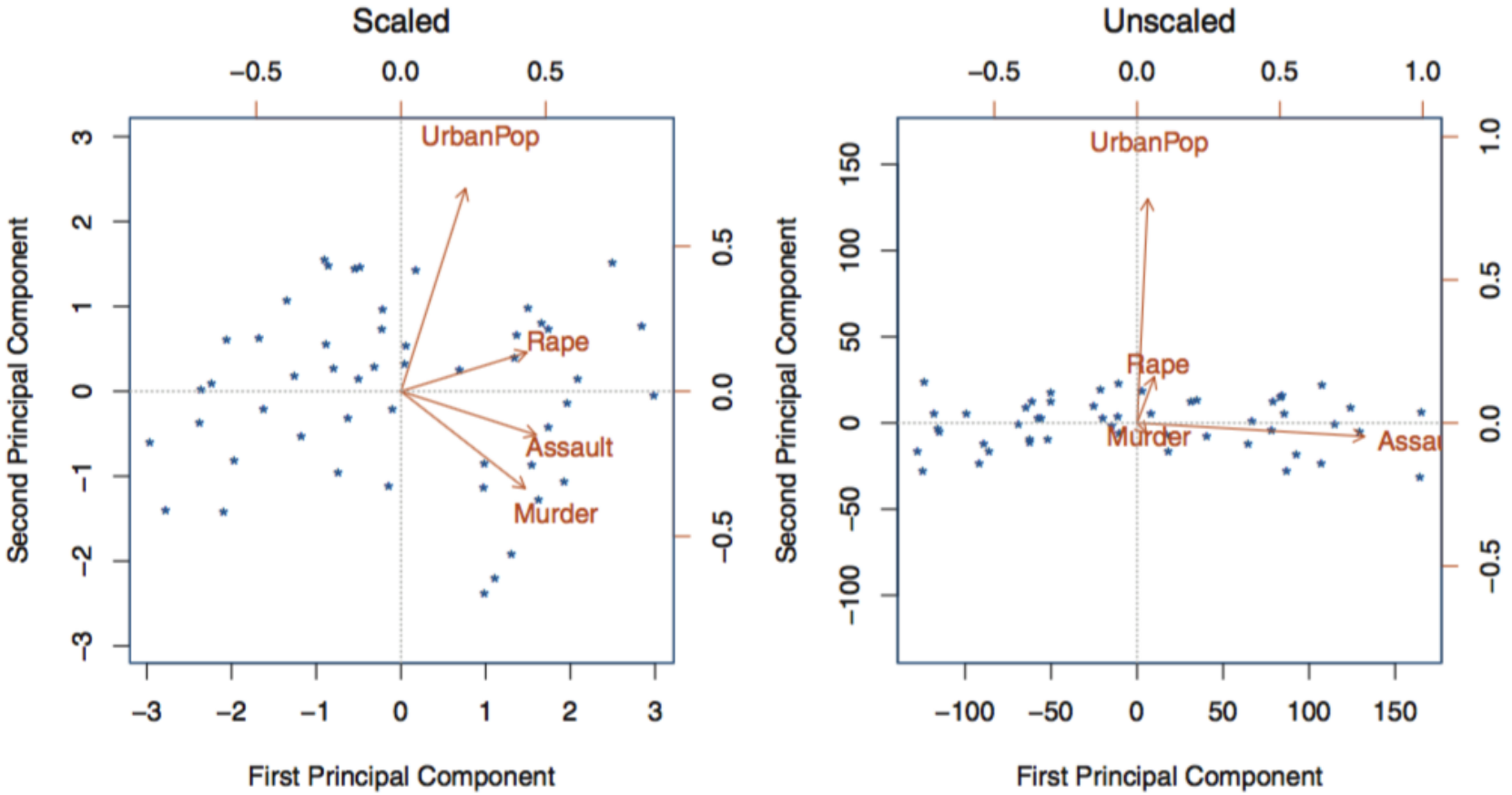
Then, the first M principal components provide the best M-dimensional approximation (in terms of Euclidean distance) to the ith observation x_{ij} :

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\Phi_{jm} \text{ (if } \mathbf{X} \text{ has 0 mean)}$$

M principal component score vectors and M principal component loading vectors can give a good approximation to the data when M is sufficiently large. If $M = p$ the approximation is exact (if $p < n$).

More on PCA: Scaling 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, it is not necessary to scale the variables.



$Var(\texttt{Murder}) = 18.97$, $Var(\texttt{Rape}) = 87.73$, $Var(\texttt{Assault}) = 6945.16$, $Var(\texttt{UrbanPop}) = 209.5$

Uniqueness of the Principal Components

- Each principal component loading vector is **unique**, up to a sign flip.
- Similarly, the score vectors are unique up to a sign change,
- It is worth noting that in the approximation formula of x_{ij}, we multiply z_{im} by Φ_{jm}. Hence, if the sign is flipped on both vectors, the final product of the two quantities is unchanged.

Proportion of Variance Explained (PVE)

How much information is lost by projecting the observations onto the first few principal components?

- We are interested in knowing the **proportion of variance explained (PVE)** by each principal component.
- The *total variance* present in a data set (assuming that the variables have been centered to have mean zero) is

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

and the variance explained by the mth principal component is

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

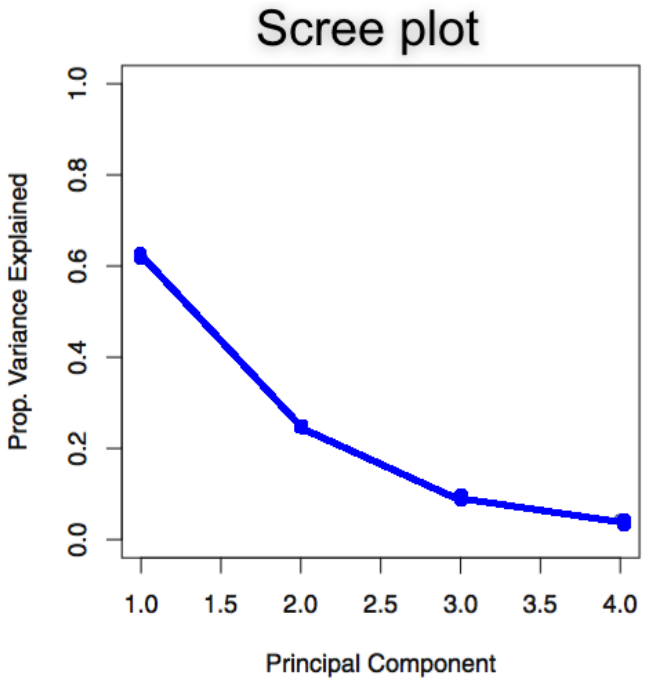
- It can be shown that $\sum_{j=1}^{p} Var(X_j) = \sum_{m=1}^{p} Var(Z_m)$.

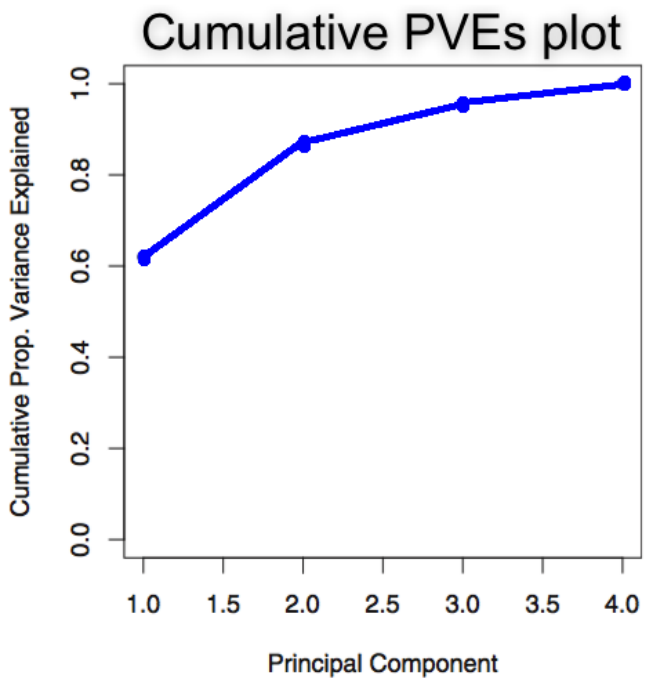
PVE and Scree plot

- the *PVE* of the *m*th principal component is given by the positive quantity between 0 and 1:

$$PVE_m = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2} = \frac{\lambda_m}{\sum_{\ell}^{p} \lambda_\ell}$$ where λ_m is the **eigen value** associated to the *m*th eigen vector.

- The *PVE* sums to 1.





How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

- No universal answer to this question. The choice is *ad hoc* and will depend on the specific application.

How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

- No universal answer to this question. The choice is *ad hoc* and will depend on the specific application.
- For visualization 3 dimensions is the limit.

How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

- No universal answer to this question. The choice is *ad hoc* and will depend on the specific application.
- For visualization 3 dimensions is the limit.
- The "**scree plot**" can be used as a guide: look for an *elbow*.

How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

- No universal answer to this question. The choice is *ad hoc* and will depend on the specific application.
- For visualization 3 dimensions is the limit.
- The "**scree plot**" can be used as a guide: look for an *elbow*.
- If the cumulative PVE is high, the new data representation is useful.

How many principal components to choose?

In practice, keep as few as possible principal components in order to visualize or interpret the data but ...

How many components are sufficient to summary the data?

- No universal answer to this question. The choice is *ad hoc* and will depend on the specific application.
- For visualization 3 dimensions is the limit.
- The "**scree plot**" can be used as a guide: look for an *elbow*.
- If the cumulative PVE is high, the new data representation is useful.
- In practice, *PCA* is a subjective approach since it is generally used as a tool for exploratory data analysis.

Outline

References

- James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. "An Introduction to Statistical Learning with Applications in R", 2nd edition, New York : "Springer texts in statistics", 2021. Site web: `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`.
- Hastie, Trevor; Tibshirani, Robert and Friedman, Jerome (2009). "The Elements of Statistical Learning (Data Mining, Inference, and Prediction), 2nd edition". New York: "Springer texts in statistics". Site web : `https://hastie.su.domains/ElemStatLearn/`