

# RECSM Summer School: Machine Learning for Social Sciences

## Session 3.2: Principal Components Analysis

Reto Wüest

Department of Political Science and International Relations  
University of Geneva



- ① Principal Components Analysis
- ② How Are the Principal Components Determined?
- ③ Interpretation of Principal Components
- ④ More on PCA
  - Scaling the Variables
  - Uniqueness of the Principal Components
  - The Proportion of Variance Explained
  - How Many Principal Components Should We Use?

# Principal Components Analysis

# Principal Components Analysis

- Suppose that we wish to visualize  $n$  observations with measurements on a set of  $p$  features,  $X_1, X_2, \dots, X_p$ , as part of an exploratory data analysis.
- How can we achieve this goal?
- We could examine two-dimensional scatterplots of the data, each of which containing two features  $(X_j, X_k) \in X, j \neq k$ .

# Principal Components Analysis

- However, there would be  $\binom{p}{2} = p(p-1)/2$  such scatterplots (e.g., with  $p = 10$  there would be 45 scatterplots).
- Moreover, these scatterplots would not be informative since each would contain only a small fraction of the total information present in the data set.
- Clearly, a better method is required to visualize the  $n$  observations when  $p$  is large.

# Principal Components Analysis

- Our goal is to find a **low-dimensional** representation of the data that captures **as much of the information as possible**. (E.g., if we can find a two-dimensional representation of the data that captures most of the information, then we can plot the observations in this two-dimensional space.)
- PCA is a method that allows us to do just this.
- It finds a low-dimensional representation of a data set that contains **as much as possible of the variation** in the data.

# Principal Components Analysis

The **idea** behind PCA is the following:

- Each of the  $n$  observations lives in a  **$p$ -dimensional space**.
- However, not all of these  $p$  dimensions are equally interesting.
- PCA seeks a small number of dimensions that are **as interesting as possible**.
- “Interesting” is determined by the amount that the observations **vary along a dimension**.
- Each of the dimensions, or **principal components**, found by PCA is a **linear combination** of the  $p$  features.

# How Are the Principal Components Determined?



## How Are the Principal Components Determined?

- The **first principal component** of features  $X_1, X_2, \dots, X_p$  is the normalized linear combination

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p \quad (3.2.1)$$

that has the **largest variance**.

- By **normalized**, we mean that  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .
- The elements  $\phi_{11}, \dots, \phi_{p1}$  are called the **loadings** of the first principal component. Together, they make up the **principal component loading vector**,  $\phi_1 = (\phi_{11} \ \phi_{21} \ \dots \ \phi_{p1})^T$ .

## How Are the Principal Components Determined?

- Why do we constrain the loadings so that their sum of squares is equal to 1?
- Without this constraint, the loadings could be arbitrarily large in absolute value, resulting in an arbitrarily large variance.
- Given an  $n \times p$  data set  $\mathbf{X}$ , how do we compute the first principal component?
- As we are only interested in variance, we **center each variable** in  $\mathbf{X}$  to have mean 0.

## How Are the Principal Components Determined?

- We then look for the **linear combination** of the feature values of the form

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip} \quad (3.2.2)$$

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

- Hence, the first principal component loading vector solves the **optimization problem**

$$\arg \max_{\phi_{11}, \dots, \phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^p \phi_{j1}^2 = 1. \quad (3.2.3)$$

## How Are the Principal Components Determined?

- Problem (3.2.3) can be solved via an **eigen decomposition** (for details, see Hastie et al. 2009, 534ff.).
- The  $z_{11}, \dots, z_{n1}$  are called the **scores** of the first principal component.
- After the first principal component  $Z_1$  of the features has been determined, we can find the second principal component  $Z_2$ .

## How Are the Principal Components Determined?

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

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}, \quad (3.2.4)$$

where  $\phi_2$  is the second principal component loading vector, with elements  $\phi_{12}, \phi_{22}, \dots, \phi_{p2}$ .

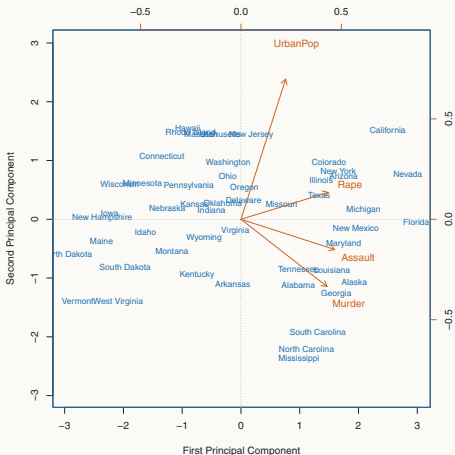
- It turns out that constraining  $Z_2$  to be **uncorrelated** with  $Z_1$  is **equivalent** to constraining the direction  $\phi_2$  to be **orthogonal** to the direction  $\phi_1$ .

## Example: USA Arrests Data

- For each of the 50 US states, the data set contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape.
- We also have for each state the population living in urban areas: UrbanPop.
- 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 0 and standard deviation 1.

# Example: USA Arrests Data

Biplot (principal component scores and loading vectors for the first two principal components)



(Source: James et al. 2013, 378)

## Example: USA Arrests Data

- In the figure, the blue state names represent the scores for the first two principal components (axes on the bottom and left).
- The orange arrows indicate the first two principal component loading vectors (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 component 0.17 (the word Rape in the plot is centered at the point  $(0.54, 0.17)$ ).



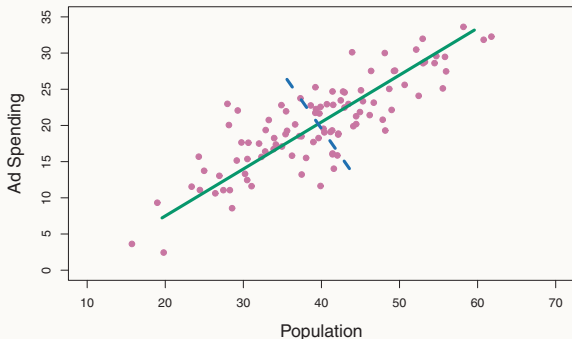
## Example: USA Arrests Data

- The first loading vector places approximately equal weight on the crime-related variables, with much less weight on `UrbanPop` (see axis on the top).  
→ Hence, this component roughly corresponds to a measure of overall crime rates.
- The second loading vector places most of its weight on `UrbanPop` and much less weight on the other three features (see axis on the right).  
→ Hence, this component roughly corresponds to the level of urbanization of a state.

# Interpretation of Principal Components

**Interpretation I:** Principal component loading vectors are the **directions** in feature space along which the **data vary the most**.

Two-dimensional data set: population size (in 10,000) and ad spending for a company (in 1,000)

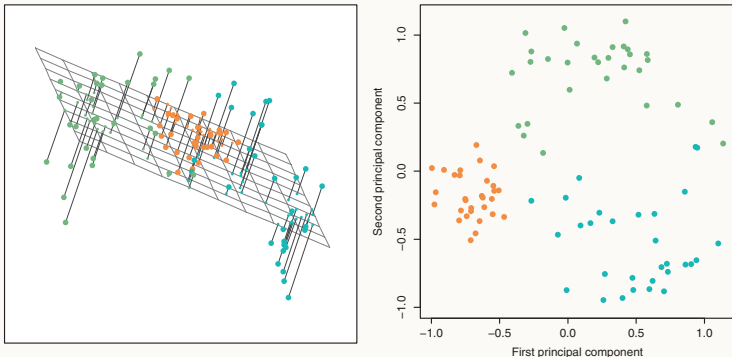


(Source: James et al. 2013, 230)

# Interpretation of Principal Components

**Interpretation II:** The first  $M$  principal component loading vectors span the  $M$ -dimensional **hyperplane** that is **closest** to the  $n$  observations.

Simulated three-dimensional data set



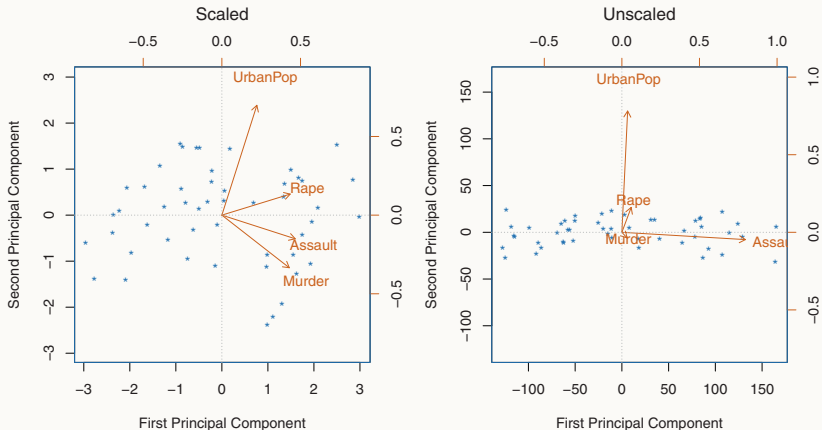
(Left: the first two principal component directions span the plane that best fits the data. Right: Projection of the observations onto the plane; the variance on the plane is maximized. Source: James et al. 2013, 380)

## Scaling the Variables

- The results obtained by PCA depend on the **scales** of the variables.
- In the US Arrests data, the variables are measured in different units: Murder, Rape, and Assault are occurrences per 100,000 people and UrbanPop is the percentage of a state's population that lives in an urban area.
- These variables have variance 18.97, 87.73, 6945.16, and 209.5, respectively.
- If we perform PCA on the **unscaled** variables, then the first principal component loading vector will have a **very large loading** for Assault.

# Scaling the Variables

## US Arrests data



(Source: James et al. 2013, 381)

## Scaling the Variables

- Suppose that Assault were measured in occurrences per 100 people rather than per 100,000 people.
- In this case, the variance of the variable would be tiny, and so the first principal component loading vector would have a **very small value** for that variable.
- We typically **scale each variable** to have a standard deviation of 1 before we perform PCA, so that the principal components do **not depend** on the choice of scaling.
- However, if the variables are measured in the **same units**, we might choose **not to scale** the variables.

## Uniqueness of the Principal Components

- Each principal component loading vector is **unique**, up to a **sign flip**.
- The reason is that a principal component loading vector specifies a direction in  $p$ -dimensional space. Flipping the sign has no effect as the direction does not change.
- Similarly, the score vectors are **unique** up to a **sign flip**, since the variance of  $Z$  is the same as the variance of  $-Z$ .

## The Proportion of Variance Explained

- Above, we performed PCA on a simulated three-dimensional data set (left panel) and projected the data onto the first two principal component loading vectors (right panel).
- In this case, the two-dimensional representation of the three-dimensional data successfully captures the major pattern in the data.
- But how much of the **information** in a data set is **lost** by projecting the observations onto the first few principal components? Or, how much of the **variance** in the data is **not contained** in the first few principal components?



## The Proportion of Variance Explained

- To answer this question, we want to know the **proportion of variance explained** (PVE) by each principal component.
- The **total variance** present in a data set is (assuming that the variables have been centered)

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2. \quad (3.2.5)$$

- The **variance explained** by the  $m$ th principal component is

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2. \quad (3.2.6)$$

## The Proportion of Variance Explained

- Therefore, the **proportion of variance explained** (PVE) by the  $m$ th principal component is

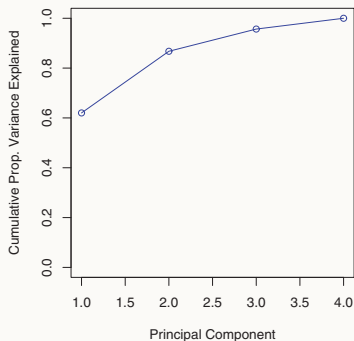
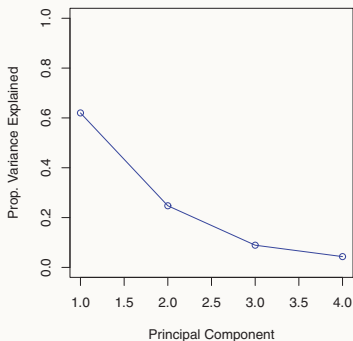
$$\frac{\sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}. \quad (3.2.7)$$

- To compute the **cumulative PVE** of the first  $M$  principal components, we can sum (3.2.7) over the first  $M$  PVEs.
- In the US Arrests data, the first principal component explains 62.0% of the variance in the data and the second principal component explains 24.7% of the variance.

# The Proportion of Variance Explained

- Together, the first two principal components explain  $\approx 87\%$  of the variance and the last two principal components explain only  $\approx 13\%$  of the variance.

PVE (scree plot) and cumulative PVE



(Source: James et al. 2013, 383)

## How Many Principal Components Should We Use?

- A  $n \times p$  data matrix  $\mathbf{X}$  has  $\min(n - 1, p)$  principal components.
- Our goal is to use the **smallest number** of principal components required to get a **good understanding** of the data.
- We typically decide on the number of principal components by examining a **scree plot** (see above).
- We do so by eyeballing the scree plot and looking for an **“elbow”** in the plot (a point at which the PVE drops off).