Universität Zürich[UZH]

Soziologisches Institut

Data Analysis – Advanced Statistics with Python

Dr. Julia Jerke

jerke@soziologie.uzh.ch

Thursday, 12.15pm – 13.45pm, AND 2.46

Session 8 – Principal component analysis

Agenda

1. Dimensionality reduction
2. Basics of principal component analysis (PCA)
3. Brief mathematical background
4. Hands on

1. Dimensionality reduction

Motivation

- So far, we focused on methods that check for a hypothesized structure in the data. Therefore, it is necessary to have an idea about the relationship between a dependent variable Y and a set of given variables $X_1, X_2, \ldots, X_p$

- We will now turn towards methods that discover structure within a presumably unstructered data set. In this case, we do not make assumptions about (directed) relationships between variables.

We will discuss two methods in this context:

1. **Cluster analysis** (this session)
2. **Principal component analysis** (next session)

Methods of dimensionality reduction

What means dimensionality reduction?

- Imagine we have p measurements $X_1, X_2, \dots, X_p$ from our observations
- The larger p is, the more complex the description, analysis, interpretation and visualization of the data will become
- However, we can make use of methods of dimensionality reduction to transform the p measurements into q features $F_1, F_2, \dots, F_q$, whereas $q \ll p$
- The aim is that the features $F_1, F_2, \dots, F_q$ retain most of the information from the original measurements $X_1, X_2, \dots, X_p$
- Common methods: Principal Component Analysis (PCA) and Factor Analysis (FA)

Dimensionality reduction

Transforming high-dimensional data into a lower dimensional space that still retains most of the basic structure, properties and information from the orginial high-dimension space

PCA versus FA

- Both methods fulfill the purpose of dimensionality reduction
- **But what is the difference between them?**

Principal component analysis	Factor analysis
Dimensionality reduction method	Dimensionality reduction method
The analysis is based on the correlations between the original variables $X_1, X_2, \ldots, X_p$	The analysis is based on the correlations between the original variables $X_1, X_2, \ldots, X_p$
Main purpose: dimensionality reduction (e.g. for subsequent analyses), no statistical model!	Main purpose: identifying latent variables (factors) that underlie and influence $X_1, X_2, \ldots, X_p$
The method indentifies **principal components**	The method identifies **factors**
The principal components are linear combinations of the original variables	The original variables are linear combinations of the underlying factors

2. Basics of principal component analysis (PCA)

Dimensionality reduction with the PCA

Motivation

- In order to simplify subsequent analyses or data visualization, we aim to reduce the number of variables
- At the same time we want to retain as much information as possible
- We, therefore, need to find a way to summarize p variables $X_1, X_2, \ldots, X_p$ into q components $PC_1, PC_2, \ldots, PC_q$ in such way that the q components carry a comparable amount of information as the p variables

Basic idea

- The PCA is based on the correlation or the covariance matrix of the variables $X_1, X_2, \ldots, X_p$
- Broadly speaking: *information = variation* and *covariation = joint information*
- Correlation of two variables means that parts of the information of one variable are also contained in the other variable, and vice versa
- The PCA makes use of this joint information which helps reducing the p variables to q components
- The resulting components are orthogonal to each other, meaning they are uncorrelated!

Empirical example

- The data reflect the US crime statistics of the year 1980 for 50 US states
- The objective is to describe the states with respect to their level of crime
- However, since we have seven different variables, comparing the states with respect to all of them will become rather confusing
- *But*: many variables correlate quite high with each other and that can be exploited for a dimensionality reduction
- The PCA will transform the seven correlated variables into $q < 7$ components that are uncorrelated

	state	label	robbery	burglary	theft	car_theft	murder	rape	assault
0	Alabama	AL	132	1526	2642	316	13	30	273
1	Alaska	AK	90	1385	3727	617	9	62	317
2	Arizona	AZ	193	2155	4891	473	10	45	401
3	Arkansas	AR	80	1119	2169	187	9	26	218
4	California	CA	384	2316	3880	742	14	58	436

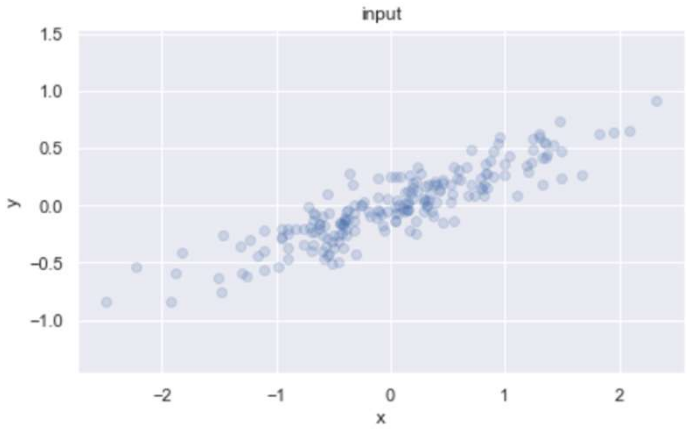
	robbery	burglary	theft	car_theft	murder	rape	assault
robbery	1.000	0.745	0.327	0.667	0.596	0.539	0.523
burglary	0.745	1.000	0.660	0.649	0.575	0.771	0.675
theft	0.327	0.660	1.000	0.377	0.119	0.614	0.390
car_theft	0.667	0.649	0.377	1.000	0.254	0.417	0.475
murder	0.596	0.575	0.119	0.254	1.000	0.671	0.604
rape	0.539	0.771	0.614	0.417	0.671	1.000	0.639
assault	0.523	0.675	0.390	0.475	0.604	0.639	1.000

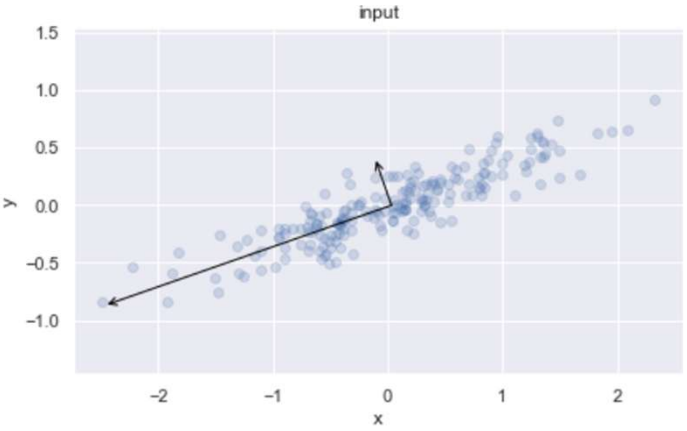
Visual illustration in the case of $p = 2$

The example is based on source code provided by Jake VanderPlas («Python Data Science Handbook»)

- Basis: we have measured the variables X_1 and X_2
- In the example data the correlation between the two variables is 0.89
- Both variables share a large proportion of information
- Reduction to one dimension, i.e. *one principal component*, is therefore possible and useful
- How can that be achieved?
  - geometrically, the determination of the principal components can be interpreted as a rotation of the axes of the coordinate system around the origin
  - The rotation is done in such way that the point cloud has a maximum variance along the first axis of the rotated coordinate system

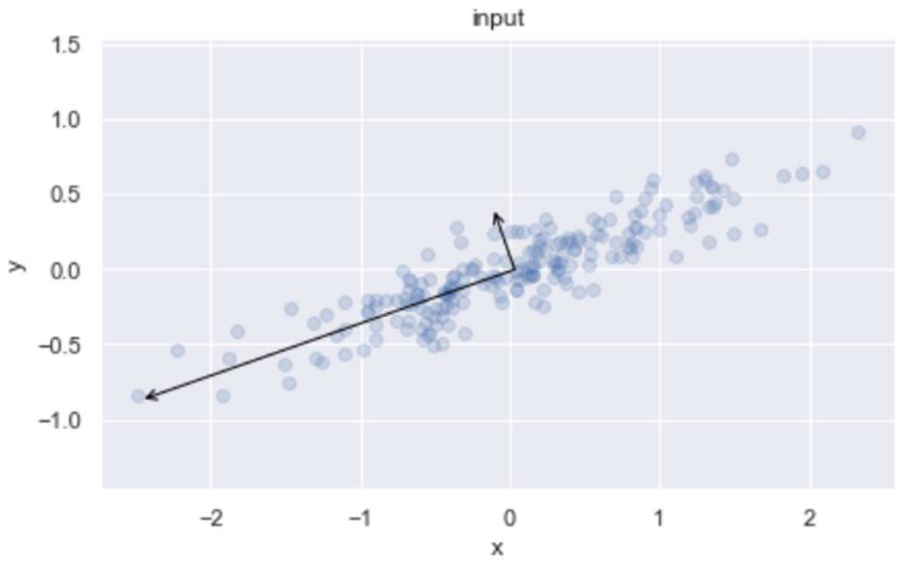
[In the case of $p > 2$: the second axis, which is orthogonal to the first axis, is also rotated in the direction of maximum variance; then the third and so on… only the last axis is fixed since it must be orthogonal to all previous axes]





Visual illustration in the case of $p = 2$

The example is based on source code provided by Jake VanderPlas («Python Data Science Handbook»)



- The **direction of the vectors** represent the principal axes
- The **length of a vector** contains the information of how much variance the respective axis explains
- To be more accurate: it represents the variance of the data when projected on that specific axis
- It also tells us how important that axis is for the description of the distribution of X_1 and X_2
- For the purpose of dimensionality reduction, we will then discard the information along the least important axes

Visual illustration in the case of $p = 2$

The example is based on source code provided by Jake VanderPlas («Python Data Science Handbook»)

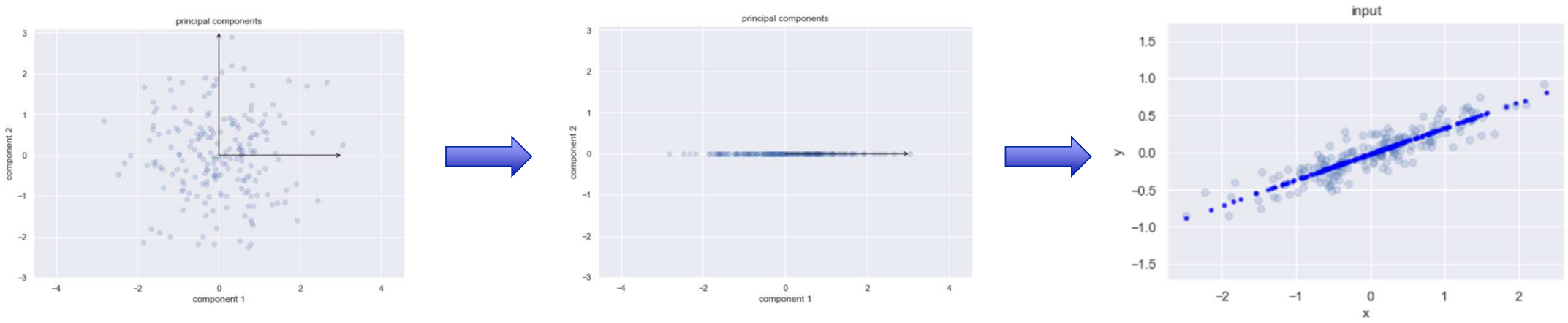
- The projection of each data point onto the principal axes build the **principal components** of the data
- The relative location of the points to each other persists
- However, the new components are now orthogonal to each other, i.e. they are uncorrelated

Visual illustration in the case of $p = 2$

The example is based on source code provided by Jake VanderPlas («Python Data Science Handbook»)

Dimensionality reduction

- Since we basically rotate the coordinate system, the PCA will in theory result in as many principal components as there are variables
- However, in practice, we will only pick the first q components and discard the remaining components
- In the example we will keep the first component and discard the second component
- As can be seen, the lower-dimensional projection still preserves a large proportion of the original variation (in the example the first component explains 97.6% of the variance)



3. Brief mathematical background of PCA

Point of departure

- We have observed p variables $X_1, X_2, \ldots, X_p$ with variances $Var(X_1), Var(X_2), \ldots, Var(X_p)$
- The total variance of the variables is given by $\mathrm{V}ar(\boldsymbol{X}) = Var(X_1) + Var(X_2) + \ldots + Var(X_p)$
- The goal of the PCA is to reduce the dimension of the data while still retaining a significant amount of the total variance of the data
- **Prerequisites of the PCA:**
  - The variables $X_1, X_2, \ldots, X_p$ must be correlated, hence share some information, otherwise the results from a PCA will not be useful
  - The stronger that the variables correlate the more information can be preserved by the dimensionality reduction
- Basis of the PCA is the **correlation matrix** in the case of standardized data or the **covariance matrix** in the case of centralized data
- *Note*:
  - The variables must either be centralized or standardized
  - If the variables have been measured on similar or at least comparable scales, centralization might be sufficient, otherwise they should be standardized

Mathematical derivation of the principal components

- The principal components are the results of an eigen value decomposition of the correlation matrix $\boldsymbol{X}_{corr}$ (for standardized data)
- Each principal component PC_j corresponds to an eigen value λ_j of the correlation matrix and its respective eigen vector $v_j = (a_{1j}, a_{2j}, \dots, a_{pj})$
- The principal components PC_j $(j = 1, \dots, p)$ can be written as linear combinations of the original variables with the eigen vectors as weights:

$$
\begin{aligned}
PC_1 &= a_{11} \cdot x_1 + a_{21} \cdot x_2 + \dots + a_{p1} \cdot x_p \\
PC_2 &= a_{12} \cdot x_1 + a_{22} \cdot x_2 + \dots + a_{p2} \cdot x_p \\
&\vdots \\
PC_p &= a_{1p} \cdot x_1 + a_{2p} \cdot x_2 + \dots + a_{pp} \cdot x_p
\end{aligned}
$$

- As said before, the whole procedure can be understood as an orthogonal rotation of the coordinate system around the origin
- The linear combinations then denote the direction of the (new) principal axes in the p-dimensional space

Side note: The eigen values are the solution of the characteristic polynom of $\boldsymbol{X}_{corr}$: $\det(\boldsymbol{X}_{corr} - \lambda \cdot I) = 0$

Mathematical derivation of the principal components

Properties

- In general, we can extract p eigen values λ_j $(j = 1, \dots, p)$
- The total variance of the principal components equals the total variance of the original variables, therefore:

$$\sum_{i=1}^{p} a_{ij}^2 = 1 \quad \text{for all } j = 1, \dots, p$$

- The principal components are pairwise uncorrelated, therefore:

$$\sum_{i=1}^{p} a_{ij} \cdot a_{ik} = 0 \quad \text{for all } j \neq k$$

- It holds that $Var(PC_j) = \lambda_j$
- We can order the eigen values in decreasing order $\lambda_1 > \lambda_2 > \dots > \lambda_p$
- The corresponding principal components are the also decreasing with respect to the explained variance
    - The first principal component PC_1 is then that linear combination of the original variables that has the highest variance
    - The second principal component PC_2 is then that linear combination of the original variables that has the second highest variance and captures a large part of the variance not explained by PC_1

Choosing an appropriate number of components

- Since we will in most cases have as many possible components as we have original variables, we need to identify criteria to choose the appropriate number of components
- Since the principal components art strictly decreasing in terms of explained variance, we usually need only the first few variables for a sufficiently good dimensionality reduction
- Common heuristics are:
  - Eigen value or Kaiser criteria: only use components that have an eigen value larger than one (these are the components that explain more variance than a single variable)
  - Keep enough components to reach a certain threshold of explained variance, e.g. 80%
  - Use the scree plot: plot the explained variance against the number of components and identify the “knee” or “elbow”
- Often a combination of the heuristics might be applied

4. Hands on

… Open *Session_8_principal_component_analysis.ipynb* in jupyter notebook