Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 5
Principal Component Analysis

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Principal component analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique used in statistics and machine learning. It transforms a large set of possibly correlated variables into a smaller set of uncorrelated variables called principal components (PCs). These components are linear combinations of the original variables. The goal is to retain as much of the original variability (information) as possible, but with fewer dimensions.

This allows maximizing the information we keep, without using variables that will cause multicolinearity, and without having to choose one variables among many. The method discovers hidden patterns and structures in the data without reference to any prior knowledge.

Why Use PCA?

- High-dimensional data is hard to analyze, visualize, and interpret.
- Variables are often correlated, which could cause redundancy.

PCA helps by:

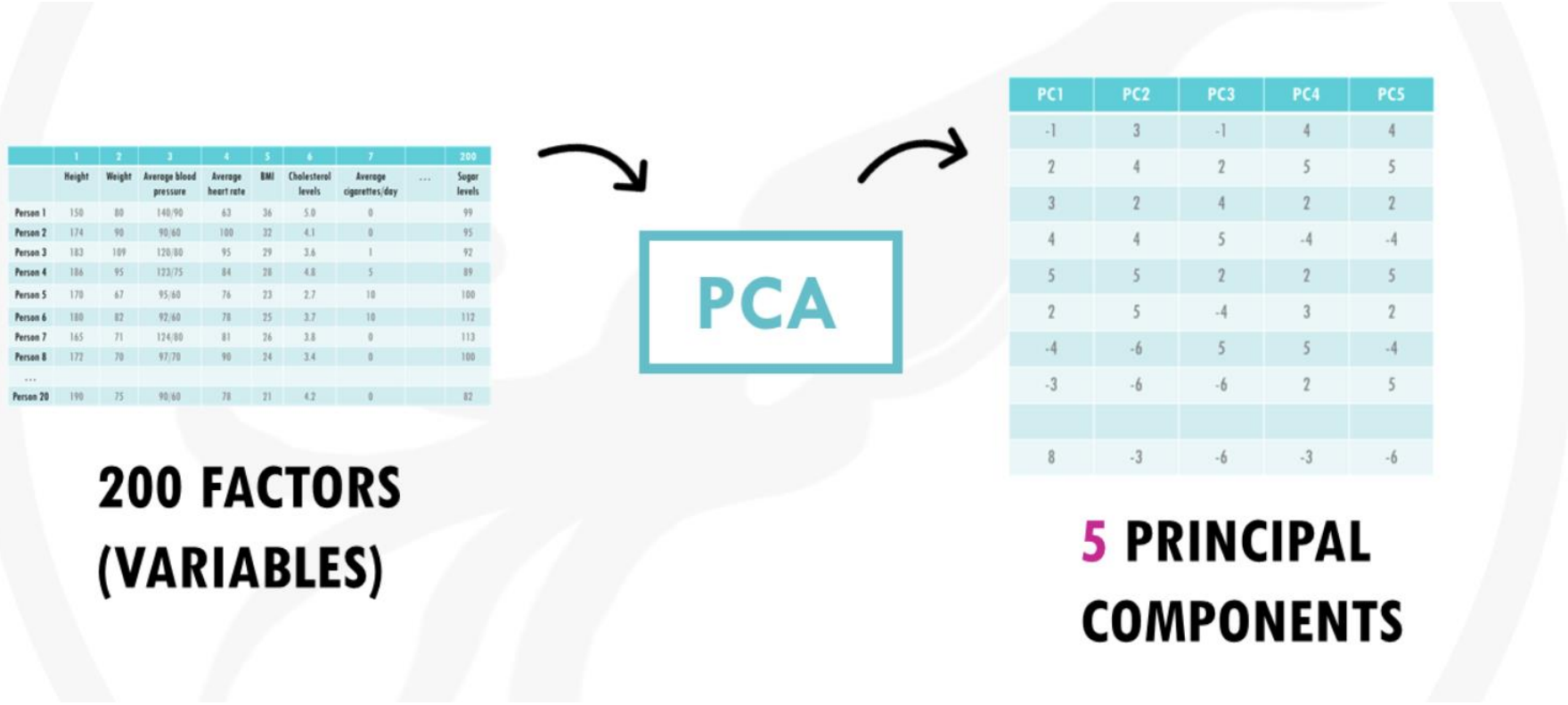
- Reducing noise and redundancy.
- Making data easier to visualize (2D or 3D plots).
- Preparing data for other algorithms (e.g., clustering, regression).

Example of PCA application

Suppose you are studying students' performance with 20 different exam scores. Many of them are correlated (math & physics, literature & history). PCA might reduce them to just 2–3 main components:

- A "STEM performance" axis.
- A "Humanities performance" axis.
- Maybe a "General test-taking ability" axis.

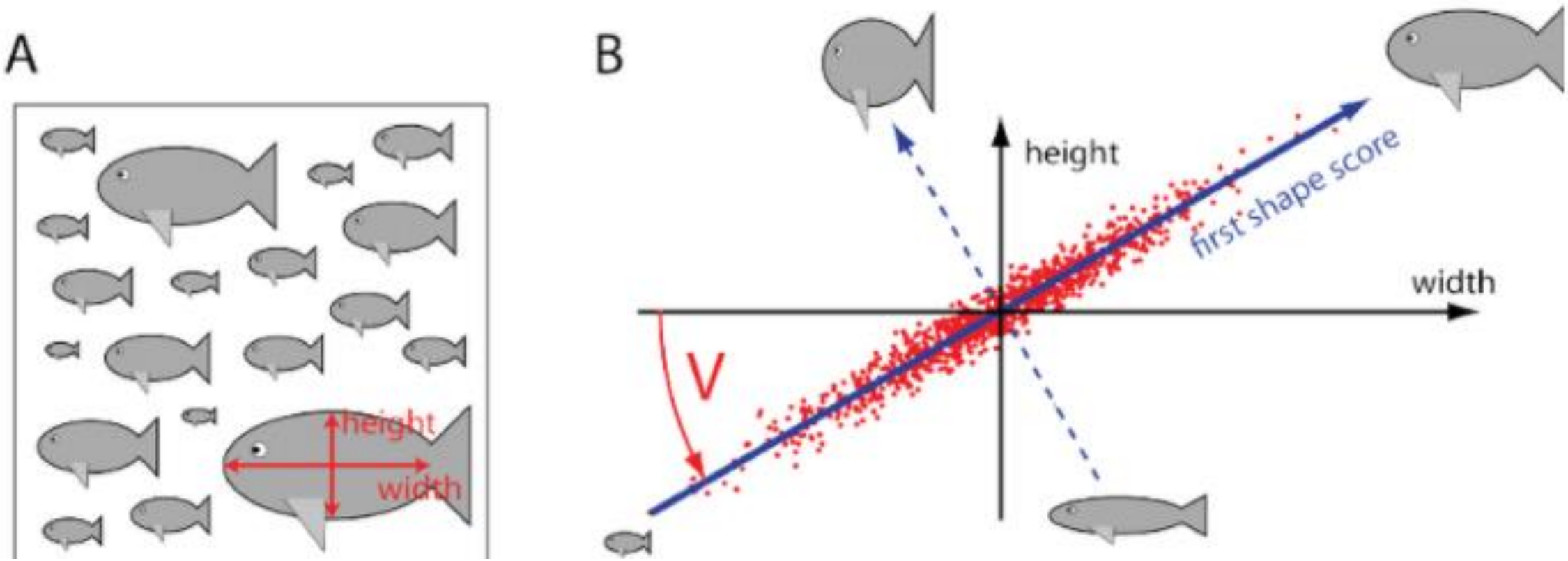
Principal component analysis



	1	2	3	4	5	6	7		200
	Height	Weight	Average blood pressure	Average heart rate	BMI	Cholesterol levels	Average cigarettes/day	...	Sugar levels
Person 1	150	80	140/90	63	36	5.0	0		99
Person 2	174	90	90/60	100	32	4.1	0		95
Person 3	183	109	120/80	95	29	3.6	1		92
Person 4	186	95	123/75	84	28	4.8	5		89
Person 5	170	67	95/60	76	23	2.7	10		100
Person 6	180	82	92/60	78	25	3.7	10		112
Person 7	165	71	124/80	81	26	3.8	0		113
Person 8	172	70	97/70	90	24	3.4	0		100
...									
Person 20	190	75	90/60	78	21	4.2	0		82

PC1	PC2	PC3	PC4	PC5
-1	3	-1	4	4
2	4	2	5	5
3	2	4	2	2
4	4	5	-4	-4
5	5	2	2	5
2	5	-4	3	2
-4	-6	5	5	-4
-3	-6	-6	2	5
8	-3	-6	-3	-6

Principal component analysis



http://setosa.io/ev/principal-component-analysis/

Principal components

Principal components (PCs) are determined in order of retaining the variation present in the original data. The first PC is chosen to account for the maximum variation. The second and subsequent PCs are chosen in a similar way to account for the remaining variance unexplained by the previous PCs. Each PC must be independent of all the previous ones. This independence or lack of correlation is referred to as orthogonality.

Principal component analysis: main steps

1. Standardization of variables' values. It will make all variables comparable since PCA is affected by scale.

2. Computation of the covariance matrix. This shows how variables vary together.

3. Computation of eigenvectors and eigenvalues of the covariance matrix to identify principal components.

4. Rank and select components (eigenvalues greater than 1).

5. Calculation of the principal components based on the values of the initial variables.

Principal component analysis: step 1

Standardization of variables' values (z-standardization).

$$z = \frac{value - mean}{standard\ deviation}$$

Principal component analysis: step 2

Computation of the covariance matrix.

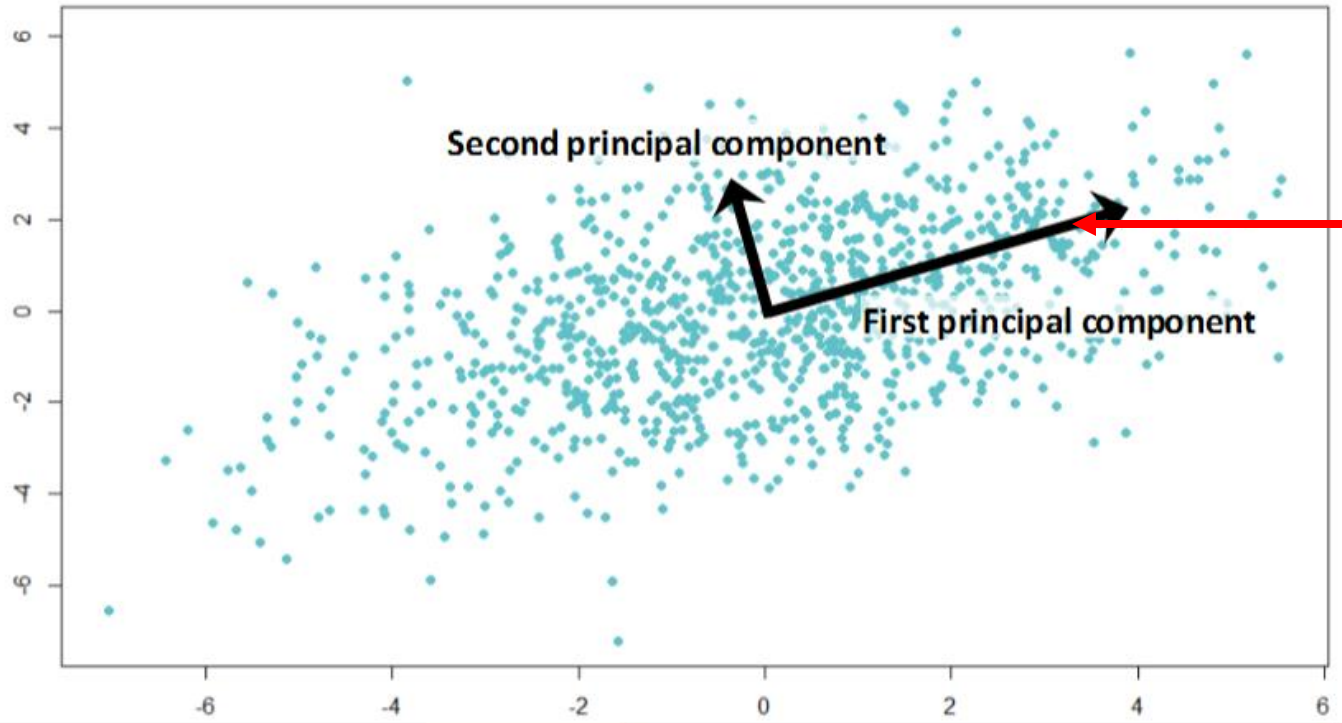
$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

Principal component analysis: step 3

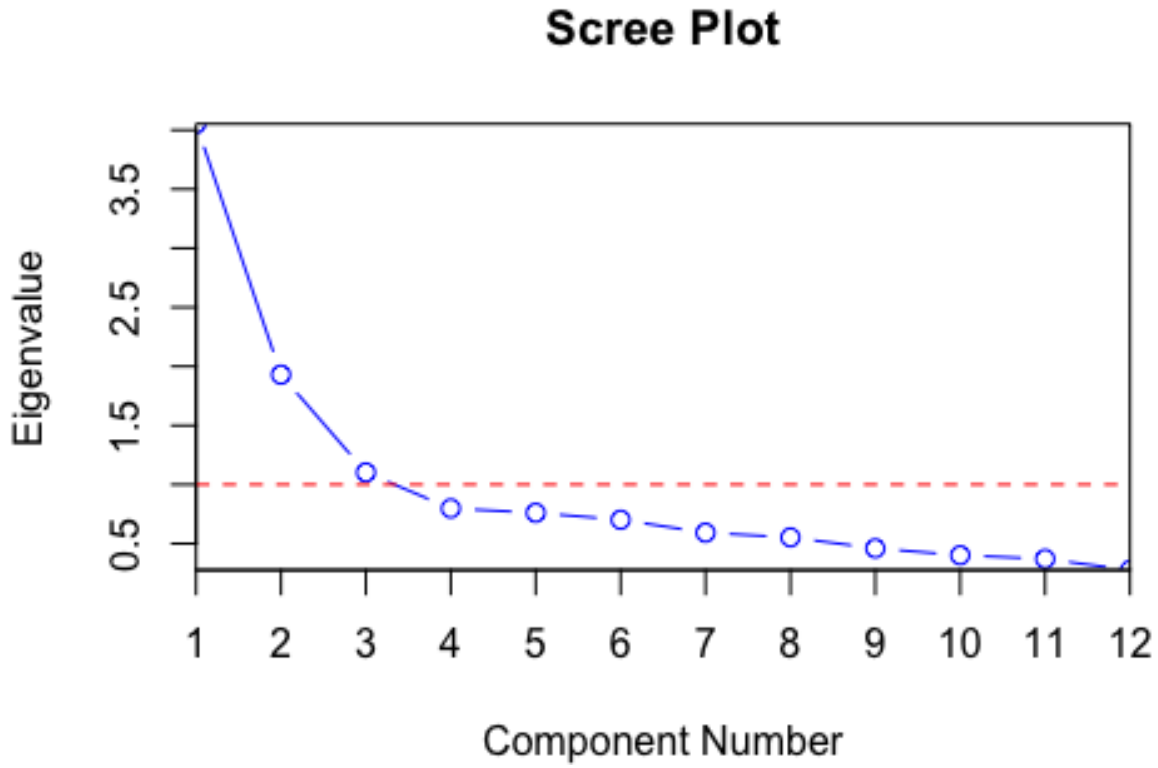
Calculation of eigenvectors and eigenvalues of the covariance matrix to identify principal components. Eigenvectors = principal components (directions of maximum variance). Eigenvalues represent the variance of the original data contained in each principal component.



We choose a line to minimize the squared deviations of the points from the line.

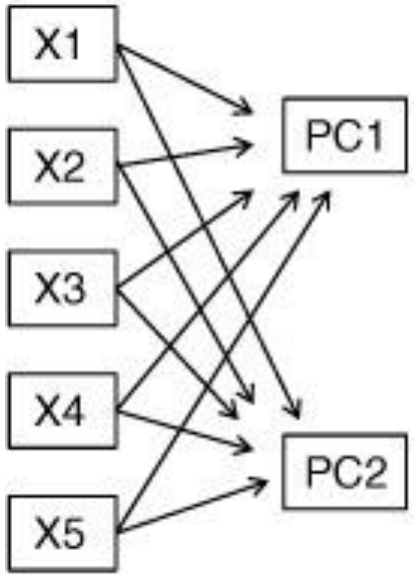
Principal component analysis: step 4

Sort eigenvalues from largest to smallest. Keep only the first few components that explain most of the variance, normally with eigenvalues greater than 1.



Principal component analysis: step 5

Calculation of the principal components based on the values of the initial variables. All initial variables are involved in the calculation of the values of the new principal components, but with different coefficients. In fact, the principal components are a linear combination of the values of the original variables.



$$PC_1 = a_1X_1 + a_2X_2 + \ldots + a_kX_k$$

$$PC_2 = b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

Selection of the number of componets

- Scree Plot: we look at the graph of eigenvalues and find an “elbow” point.
- Kaiser’s Rule: keep components with eigenvalue > 1.
- Explained Variance: keep enough PCs to capture, say, 80–90% of the variance.

Principal components' properties

- We receive as many principal components as there are initial explanatory variables.
- Each principal component is expressed as a linear combination of the original variables.
- The total variance of all input variables is equal to the total variance of all principal components. The principal components are selected so that the variance of each principal component at each step is as large as possible, so often the variance of the first principal component absorbs a significant part of the total variance of all initial variables. That is, the first several principal components can absorb a large share of the total variance of the original variables.

Example: correlation matrix

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	0.772	1.000				
Interest	0.646	0.879	1.000			
Talk 2	0.074	–0.120	0.054	1.000		
Selfish	–0.131	0.031	–0.101	0.441	1.000	
Liar	0.068	0.012	0.110	0.361	0.277	1.000

Factor 1

Factor 2

There are two groups of interrelating variables. Therefore, these variables might be measuring some common underlying dimensions.

Component 1: the better your social skills the more interesting and talkative you are likely to be.

Component 2: selfish people are likely to lie and talk about themselves.

Example

We want to measure different aspects of what might make a person popular. We will measure several aspects of person's popularity:

1) Social Skills;

2) Selfishness;

3) Interest (how interesting others find them);

4) Talk 1 (the proportion of time they spend talking about the other person during a conversation);

5) Talk 2 (the proportion of time they spend talking about themselves);

6) Liar (propensity to lie to people).

Example

Characteristics of an apartment:

- total area
- number of bus stops within walking distance
- number of living rooms
- number of schools within walking distance
- number of balconies
- number of grocery stores within walking distance

What components could be extracted?

Example

Characteristics of a restaurant:

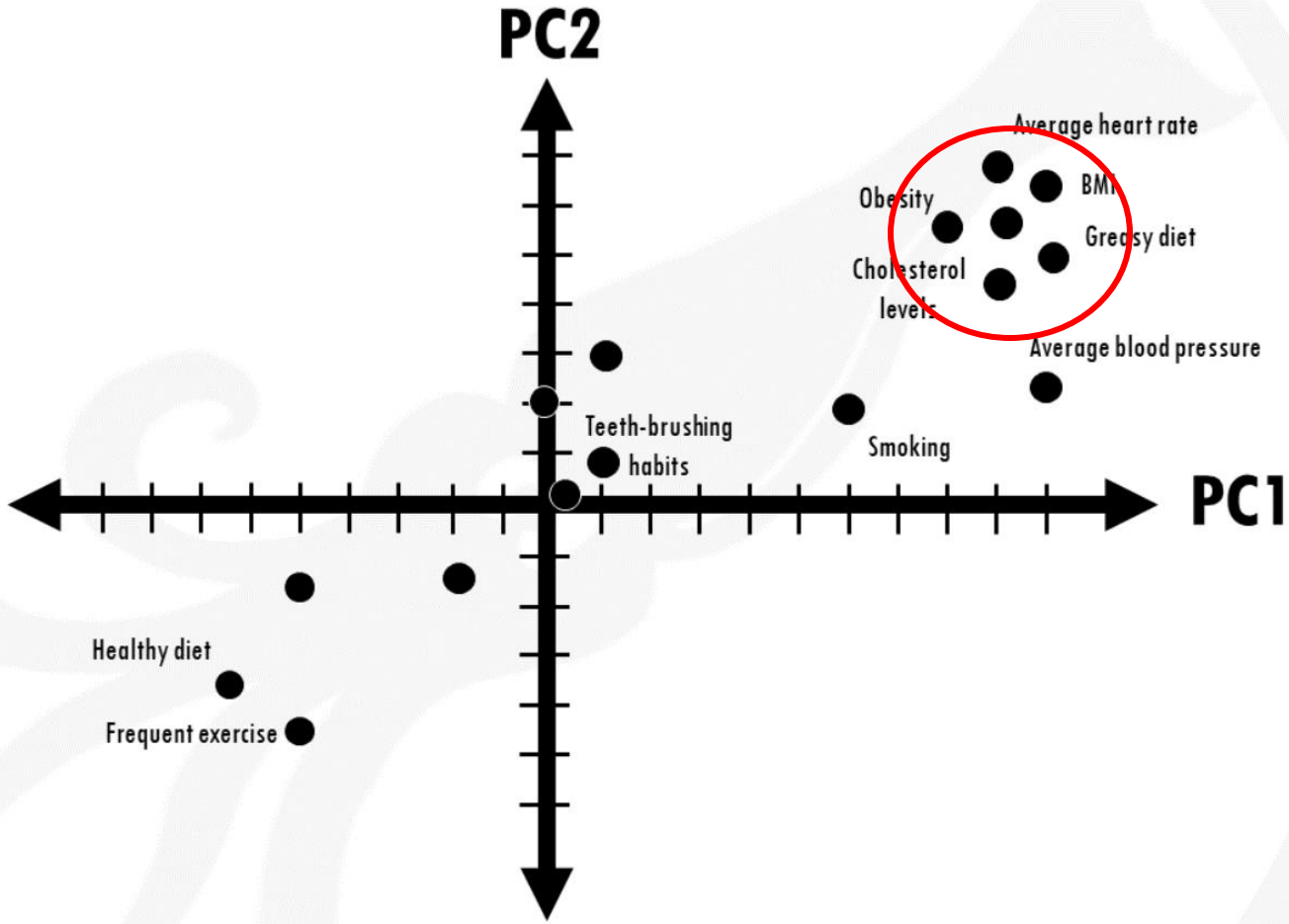
- taste of food
- food temperature
- waiting time
- cleanliness
- freshness of food
- staff behavior

What components could be extracted?

Principal Component Loadings

Principal component loadings indicate the contribution of the variables to each principal component. Each variable gets a loading, or weight, for each principal component, which tells how much it contributes to that principal component. Based on them it's possible to define which variables are **influential**, and also how the variables are **correlated**.

Visual representation of loadings



Variables that strongly correlate are grouped together. They have similar loadings.

	Cost	IT	Org
The product benefits don't outweigh the cost	0.68	0.06	-0.27
Price is prohibitive	0.49	-0.16	0.13
Overall implementation costs	0.41	0.38	0.14
We do not have sufficient technical resources	-0.21	0.58	-0.12
Our IT department cannot support your product	0	0.38	-0.25
Product is not consistent with our business strategy	0.01	-0.03	0.45
We can't reach a consensus in our organization	-0.03	-0.07	0.57
I need to develop an ROI, but cannot or have not	-0.06	0.17	0.62
Your product does not have a feature we require	-0.4	0.05	0.08
We are locked into a contact with another product	-0.21	0.08	-0.27
Other (please specify)	-0.5	-0.06	-0.1
We have no reason to switch	-0.07	-0.77	-0.25

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Factor4	Communality
Academic record	0.481	0.510	0.086	0.188	0.534
Appearance	0.140	0.730	0.319	0.175	0.685
Communication	0.203	0.280	0.802	0.181	0.795
Company Fit	0.778	0.165	0.445	0.189	0.866
Experience	0.472	0.395	-0.112	0.401	0.553
Job Fit	0.844	0.209	0.305	0.215	0.895
Letter	0.219	0.052	0.217	0.947	0.994
Likeability	0.261	0.615	0.321	0.208	0.593
Organization	0.217	0.285	0.889	0.086	0.926
Potential	0.645	0.492	0.121	0.202	0.714
Resume	0.214	0.365	0.113	0.789	0.814
Self-Confidence	0.239	0.743	0.249	0.092	0.679
Variance	2.5153	2.4880	2.0863	1.9594	9.0491
% Var	0.210	0.207	0.174	0.163	0.754

Factor-1 → Employee fit and Potential for growth in the company

Factor-2 → Personal Qualities

Factor-3 → Work Skills

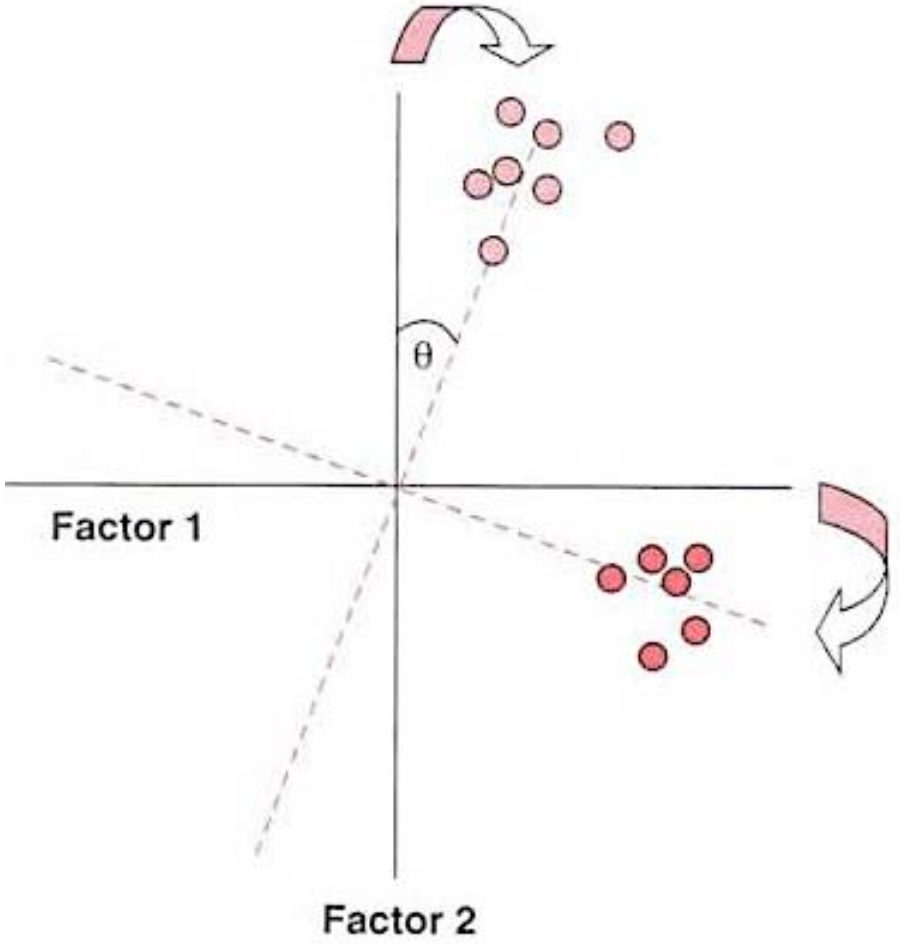
Factor-4 → Writing Skills

→ Explain 75.4% of the variati

Rotation

The rotation aims to improve the interpretability and comprehension of the principal components. We can rotate the components to create a new, more comprehensible coordinate system that is easier to read. **Varimax** rotation technique attempts to maximize the dispersion of loadings within components.

After extracting components, you can “rotate” the axes to make the loadings (the correlations between variables and components) easier to interpret.



Measure of sampling adequacy

The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy is a statistic used to assess whether PCA is suitable for a dataset. It evaluates the proportion of variance among variables that may be common and thus explainable by underlying components. KMO values range from 0 to 1: low values indicate weak or random correlations (PCA is inappropriate), while high values suggest that PCA can effectively summarize the data into reliable components.

This coefficient is typically calculated as a preliminary check before applying PCA. If the KMO value is low, we can try adjusting the set of variables or removing outliers to improve it.

KMO

$$KMO = \frac{\sum\sum_{j \neq k} r_{jk}^2}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} p_{jk}^2}$$

where r_{jk} is a pairwise correlation between variables, p_{jk} is a partial correlation.

KMO ≥ .9	marvelous
KMO ≥ .8	meritorious
KMO ≥ .7	middling
KMO ≥ .6	mediocre
KMO ≥ .5	miserable
KMO < .5	unacceptable

Advantages of PCA

- Easy to calculate and compute.
- Speeds up machine learning computing processes and algorithms.
- Prevents predictive algorithms from data overfitting issues.
- Increases performance of ML algorithms by eliminating unnecessary correlated variables.
- Helps reduce noise that cannot be ignored automatically.

Limitations of PCA

- PCA assumes that the data's underlying structure is **linear**. This means it won't be able to capture patterns if the data follows a non-linear distribution.
- PCA assumes that components with **large variances** are important. However, this might not always be the case. Sometimes, variables with smaller variance may also carry important information.
- The **output** of PCA, the principal components, are often hard to interpret. Since they're a combination of the original features, they don't carry the same interpretability as the original dataset.
- PCA can only be applied to **numeric** data.
- While reducing dimensions, PCA **discards** the components with the least variance, which could sometimes include important information.

Useful links

- https://buildmedia.readthedocs.org/media/pdf/factor-analyzer/latest/factor-analyzer.pdf
- https://devopedia.org/principal-component-analysis
- https://byjus.com/maths/eigen-values
- https://www.youtube.com/watch?v=6m83zA2yt3A
- https://www.youtube.com/watch?v=a2U4F7cxFHc

Thank you for your attention!