



POLITECNICO
MILANO 1863

Multivariate statistics for market research

AY 2024/2025 - Gloria Peggiani

Agenda

- Introduction
- Principal component analysis
- Cluster analysis

Multivariate Statistics

Multivariate statistics refers to a branch of statistics that involves the application of statistical methods that **simultaneously analyze multiple variables**. This is in contrast to univariate statistics, which involves the analysis of a single variable. In multivariate statistics, the goal is often to understand the relationships among variables, identify patterns, and make predictions.

Traditional classification of multivariate statistical methods revolves around the idea of **dependency between variables** (Kendall 1957):

Dependence multivariate methods

analyze the associations between **two sets of variables**, where one set represents the dependent variable (or variables) and the other set several independent variables.

Interdependence multivariate methods

explore the mutual association across all variables **without distinguishing** between variable types.

Multivariate Statistics

These methods can be used to either confirm a priori established theories or identify data patterns and relationships. Specifically, they are **confirmatory** when testing the hypotheses of existing theories and concepts, and **exploratory** when they search for patterns in the data in case there is no or only little prior knowledge on how the variables are related.

Common techniques used in multivariate statistics include:

	Primarily Exploratory	Primarily Confirmatory Techniques
Dependence Multivariate Methods	e.g., Partial Least Square Structural Equation Modeling ...	e.g., MANOVA, Multiple Regressions, Covariance-based Structural Equation Modeling ...
Interdependence Multivariate Methods	e.g., Cluster Analysis, RFM Analysis, Exploratory Factor Analysis, Principal Component Analysis ...	e.g., Confirmatory Factor Analysis ...

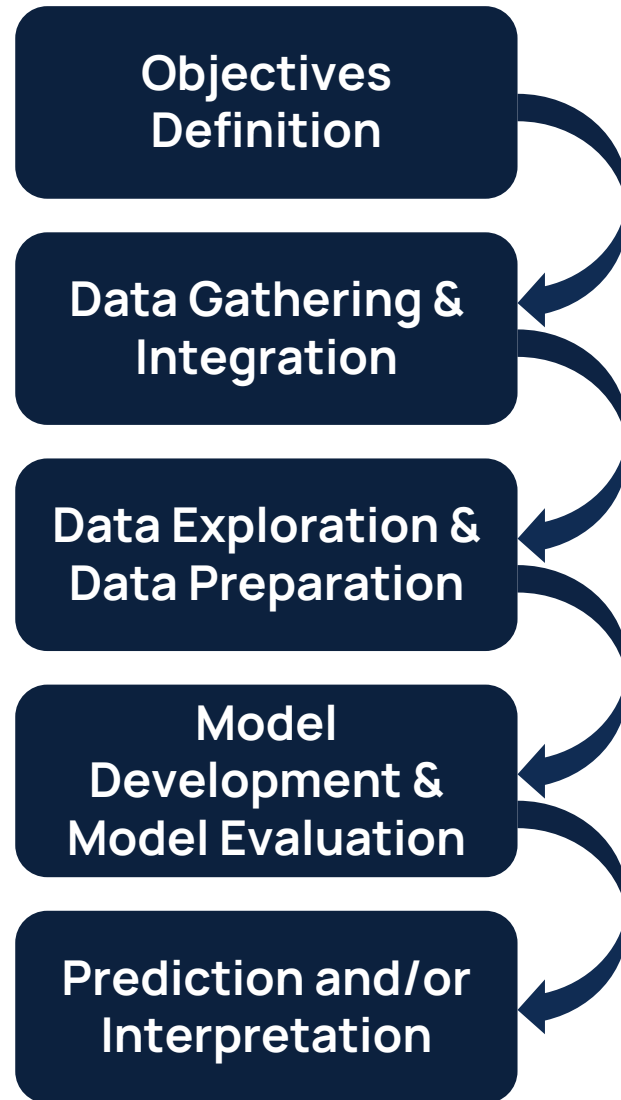
Principal Component Analysis

01

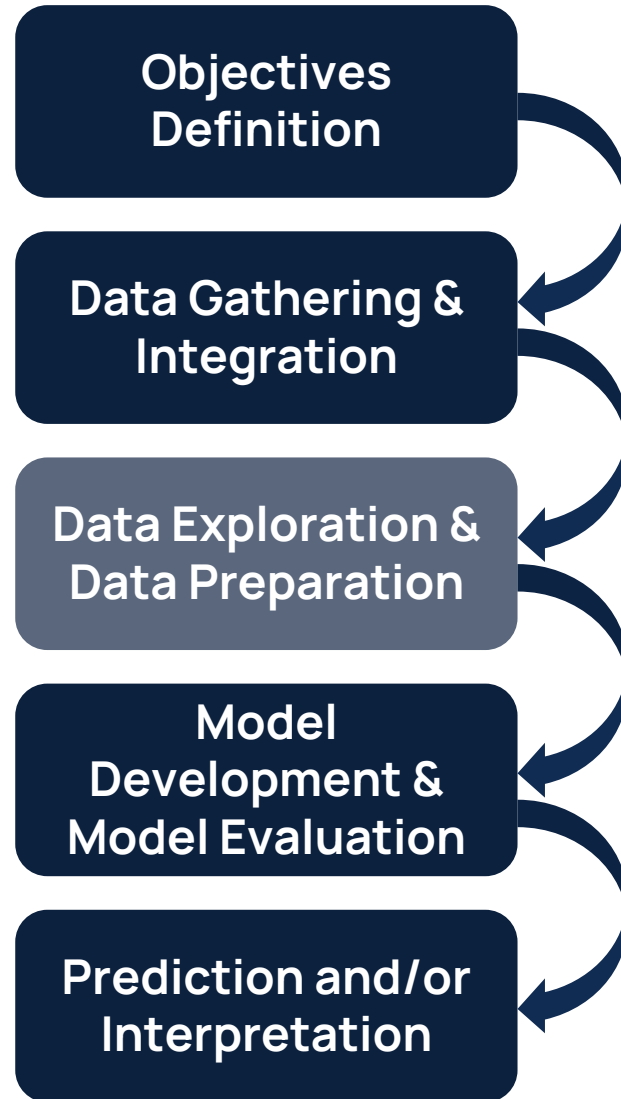
Principal Component Analysis (PCA)

- Principal Component Analysis, commonly known as PCA, is a technique for **reducing the dimensionality of large datasets**. It accomplishes this by converting a substantial set of variables into a more compact one while **retaining the majority of information** from the original dataset.
- **Important note:** principal components analysis **is conceptually different from factor analysis**. However, the two terms are often used interchangeably in several disciplines.

Data analysis process

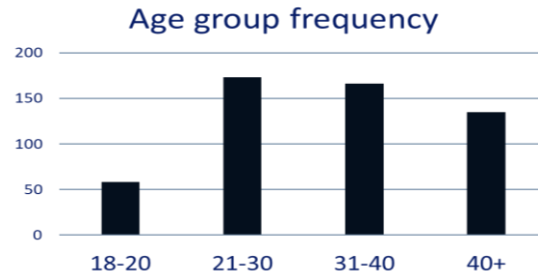


Data analysis process

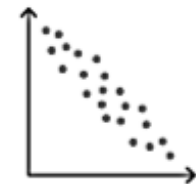
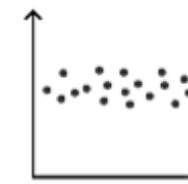
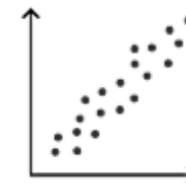
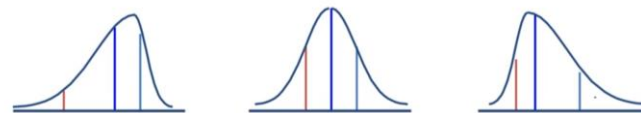
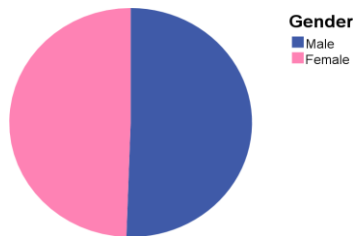
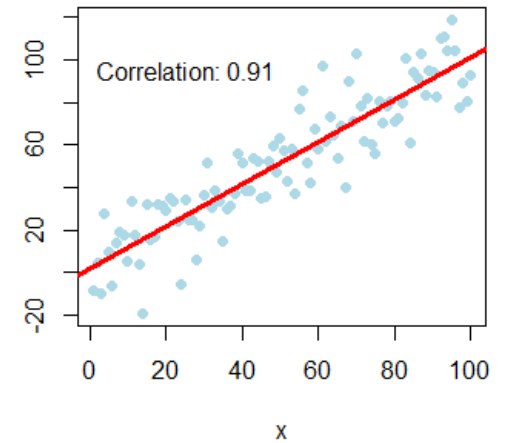
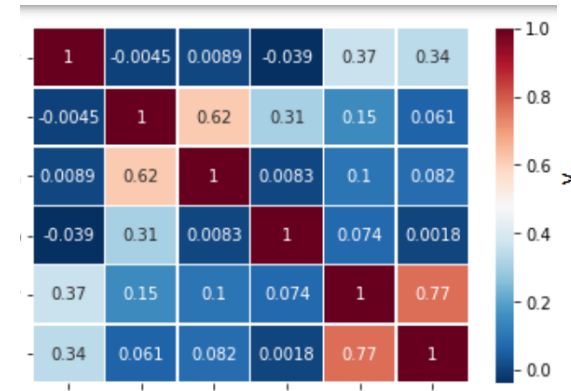


Data exploration

WillingnessToPay	
Mean	30,68364662
Standard Error	0,196944788
Median	30,6
Mode	29,6
Standard Deviation	4,542556198
Sample Variance	20,63481681
Kurtosis	25,06914004
Skewness	2,297700036
Range	61
Minimum	19
Maximum	80
Sum	16323,7
Count	532



		Count
Age group	18-20	58
	21-30	173
	31-40	166
	40+	135
Total		532



Data preparation

Data Validation

To identify and remove anomalies and inconsistencies

Data Transformation

To improve the accuracy and efficiency of algorithms

Data Size Reduction

To obtain a dataset with a lower number of attributes and records but which is as informative as the original dataset

Data preparation

Data Validation

- **Incompleteness**
 - Elimination
 - Inspection
 - Identification
 - Substitution
- **Noise**
 - Outliers detection
 - Outliers elimination
- **Inconsistencies**
- **Imbalanced data**
 - Downsampling

Data Transformation

To improve the accuracy and efficiency of algorithms

Data Size Reduction

To obtain a dataset with a lower number of attributes and records but which is as informative as the original dataset

Data preparation

Data Validation

To identify and remove anomalies and inconsistencies

Data Transformation

To improve the accuracy and efficiency of algorithms

Data Size Reduction

To obtain a dataset with a lower number of attributes and records but which is as informative as the original dataset

Data preparation

Data Validation

To identify and remove anomalies and inconsistencies

Data Transformation

- **Data Standardization**
- **Conversion of Categorical Variables**
→ Dummy Variables Creation

Data Size Reduction

To obtain a dataset with a lower number of attributes and records but which is as informative as the original dataset

Data preparation

Data Validation

To identify and remove anomalies and inconsistencies

Data Transformation

To improve the accuracy and efficiency of algorithms

Data Size Reduction

To obtain a dataset with a lower number of attributes and records but which is as informative as the original dataset

Data preparation

Data Validation

To identify and remove anomalies and inconsistencies

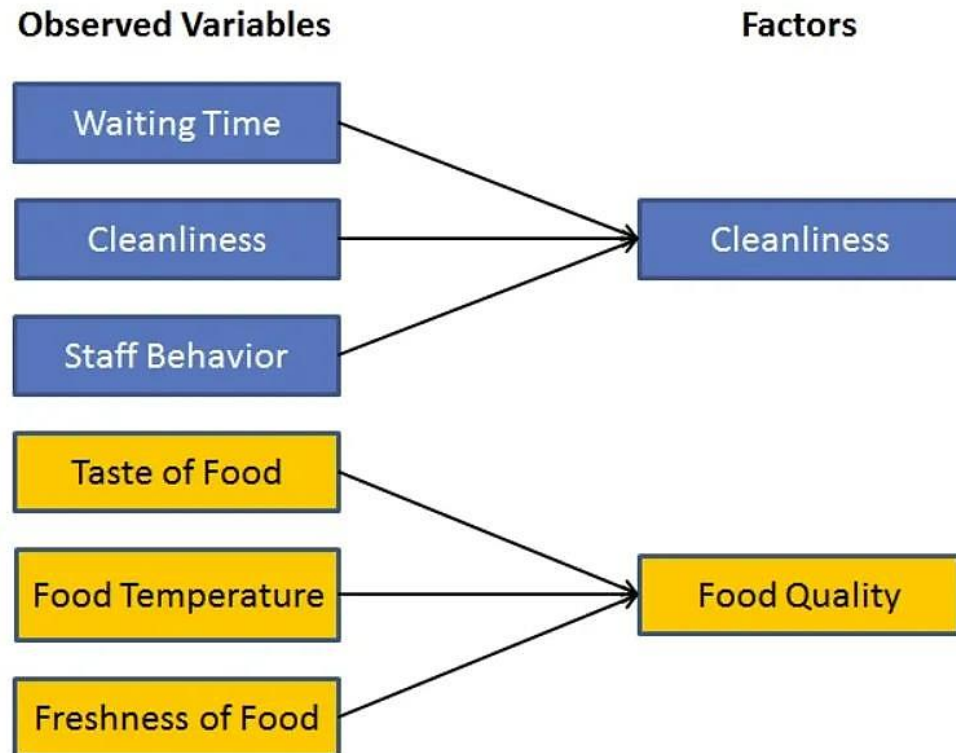
Data Transformation

To improve the accuracy and efficiency of algorithms

Data Size Reduction

- Records Reduction by Sampling
- Variables Reduction by Selection
- **Variables Reduction by Projection**
- Values Reduction by Aggregation

Data preparation



Data Size Reduction

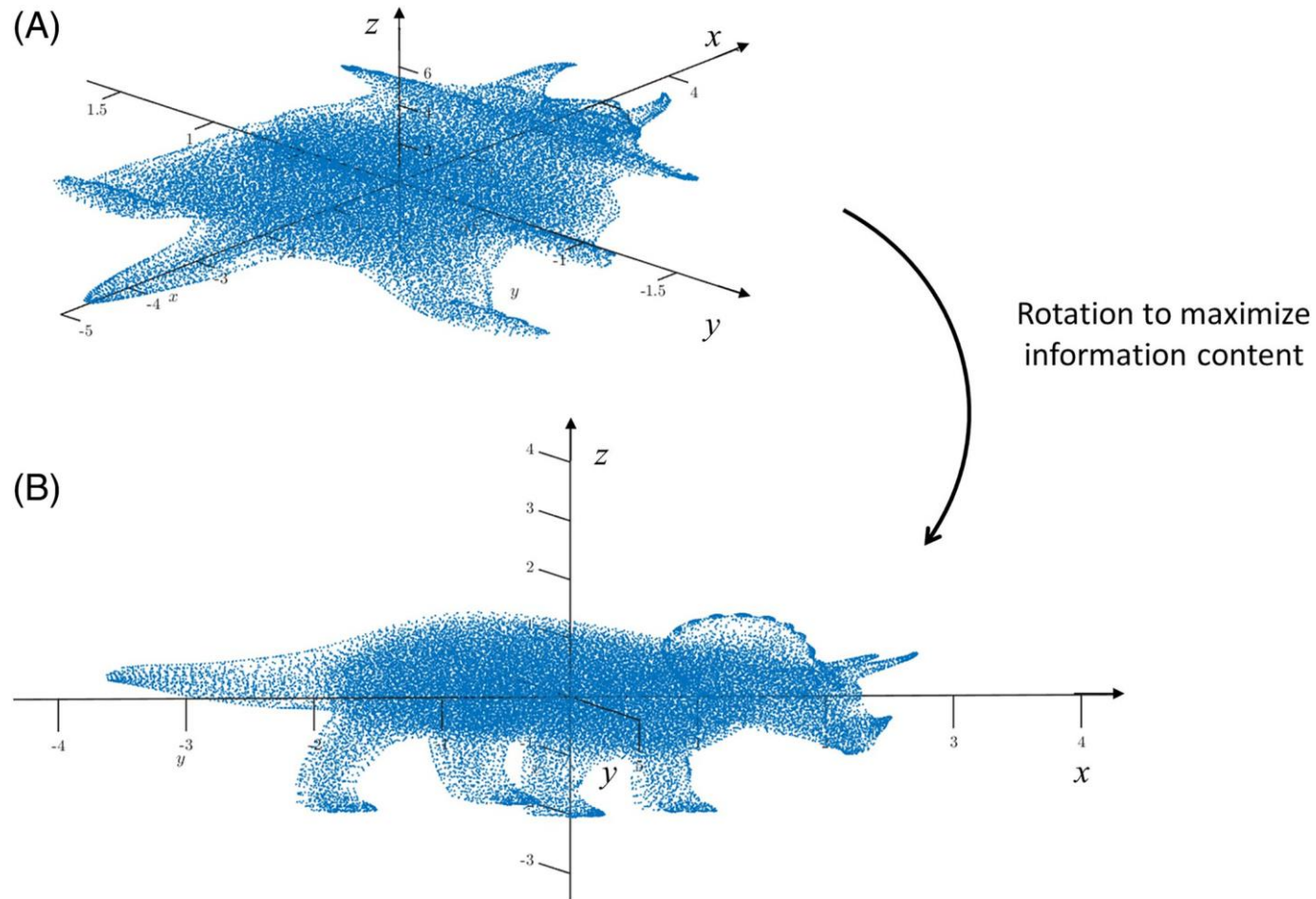
- Records Reduction by Sampling
- Variables Reduction by Selection
- **Variables Reduction by Projection**
- Values Reduction by Aggregation

When and Why is PCA needed

- **Removing superfluous/unrelated variables:** If one component only loads on one variable, this could be an indication that this variable is not related to the other variables in your data set and might not be measuring anything of importance to your particular study (i.e., it is measuring some other construct or measure).
- **Reducing redundancy in a set of variables:** If you have measured many variables and you believe that some of the variables are measuring the same underlying construct, you might have variables that are highly correlated.
- **Removing multicollinearity:** If you have two or more variables that are highly correlated, PCA can reduce the highlighted correlated variables into principal components.

PCA - Intuition

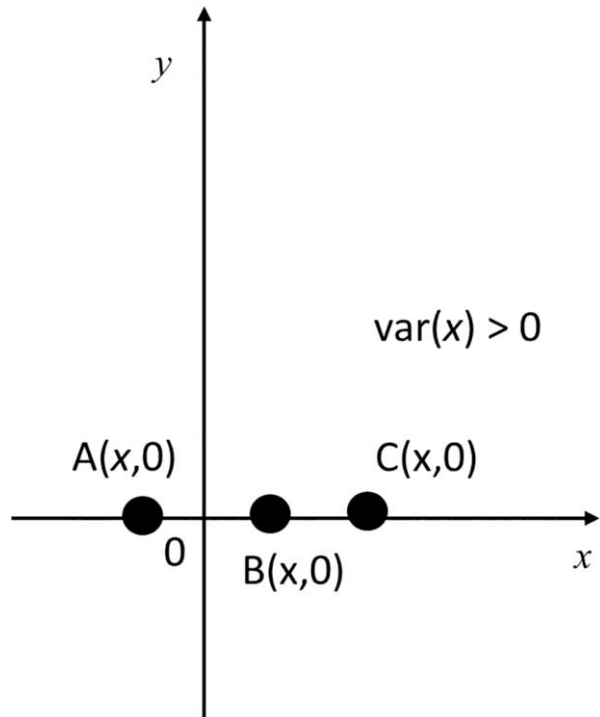
Dataset: 36,876 observations, recorded for 3 variables x , y , and z .



PCA – Intuition

INFORMATION: VARIATION (VARIABILITY)

The more spread out the data points are in a particular direction, the higher the amount of information contained in that direction.



- $A(x,y)$, $B(x,y)$, and $C(x,y)$ are scattered around $(0,0)$ and are distinguishable.
- $\text{var}(x) > 0$

PCA – Intuition

INFORMATION: VARIATION (VARIABILITY)

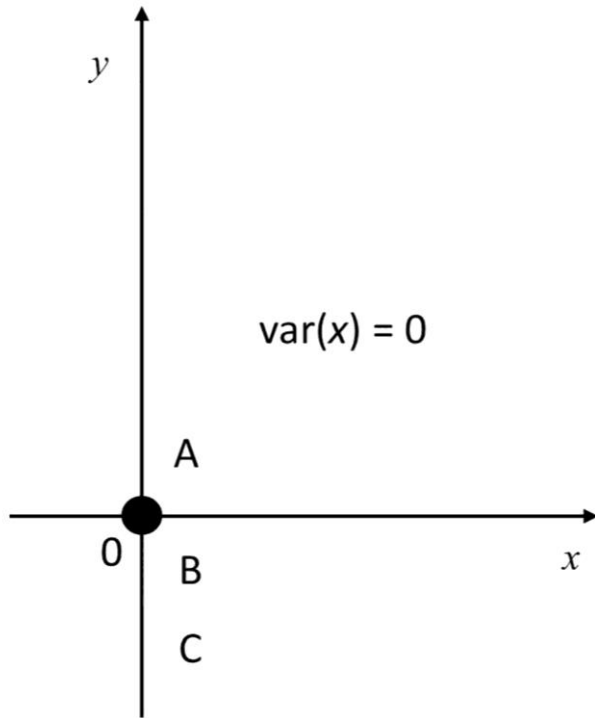
The more spread out the data points are in a particular direction, the higher the amount of information contained in that direction.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$

PCA – Intuition

INFORMATION: VARIATION (VARIABILITY)

The more spread out the data points are in a particular direction, the higher the amount of information contained in that direction.

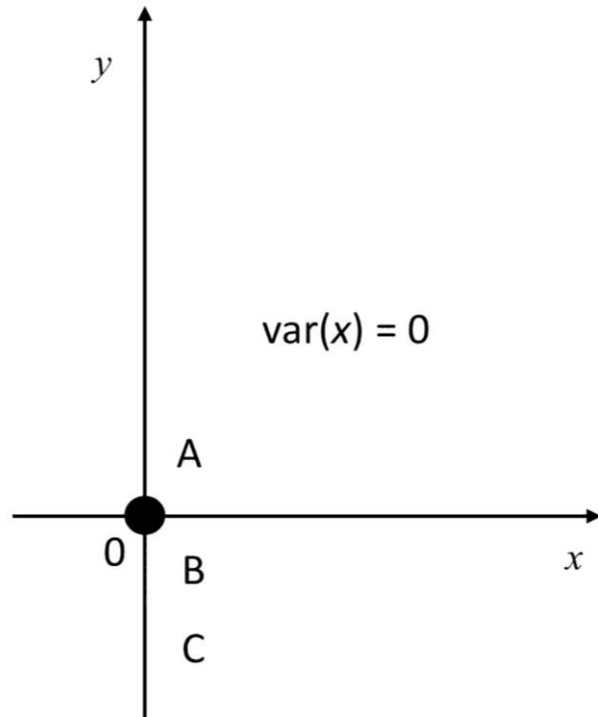


- $A(x,y)$, $B(x,y)$, and $C(x,y)$ overlap on $(0,0)$ and are indistinguishable
- No variability among the coordinates \rightarrow no information that allows to distinguish the three points.

PCA – Intuition

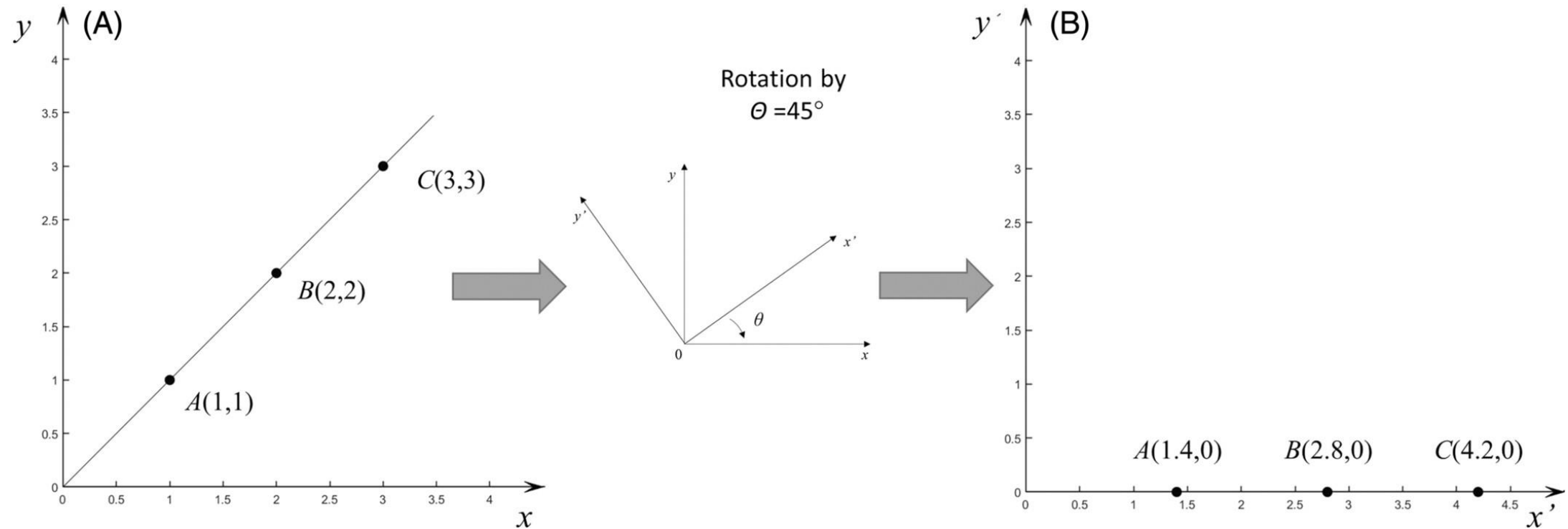
INFORMATION: VARIATION (VARIABILITY)

The more spread out the data points are in a particular direction, the higher the amount of information contained in that direction.

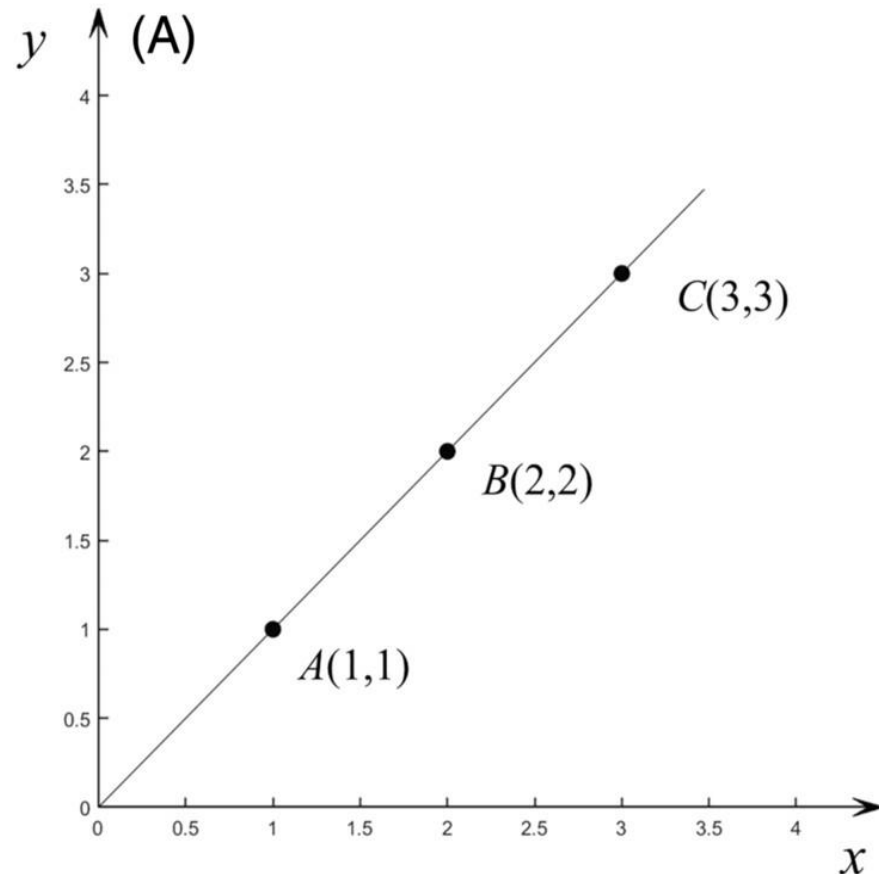


- $A(x,y)$, $B(x,y)$, and $C(x,y)$ overlap on $(0,0)$ and are indistinguishable
- No variability among the coordinates \rightarrow no information that allows to distinguish the three points.

PCA – From intuition to optimization



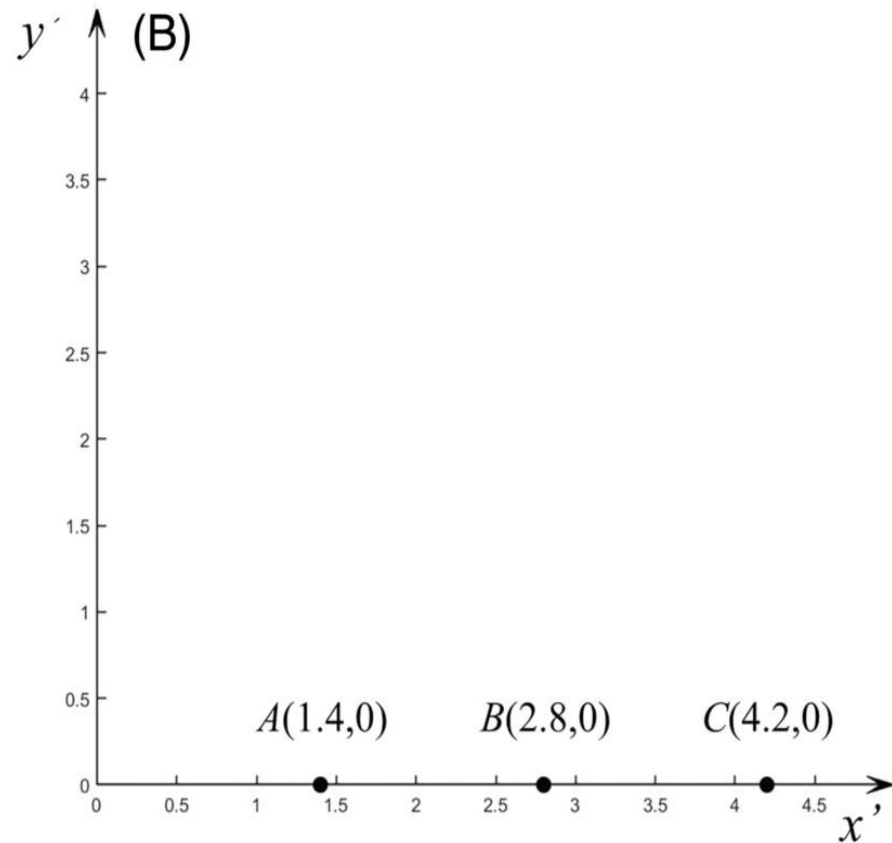
PCA – From intuition to optimization



Three points $A(1,1)$, $B(2,2)$, and $C(3,3)$ in the reference system xOy are aligned along the $\theta=45^\circ$ direction.

- $\text{var}(x)=1$
- $\text{var}(y)=1$

PCA – From intuition to optimization



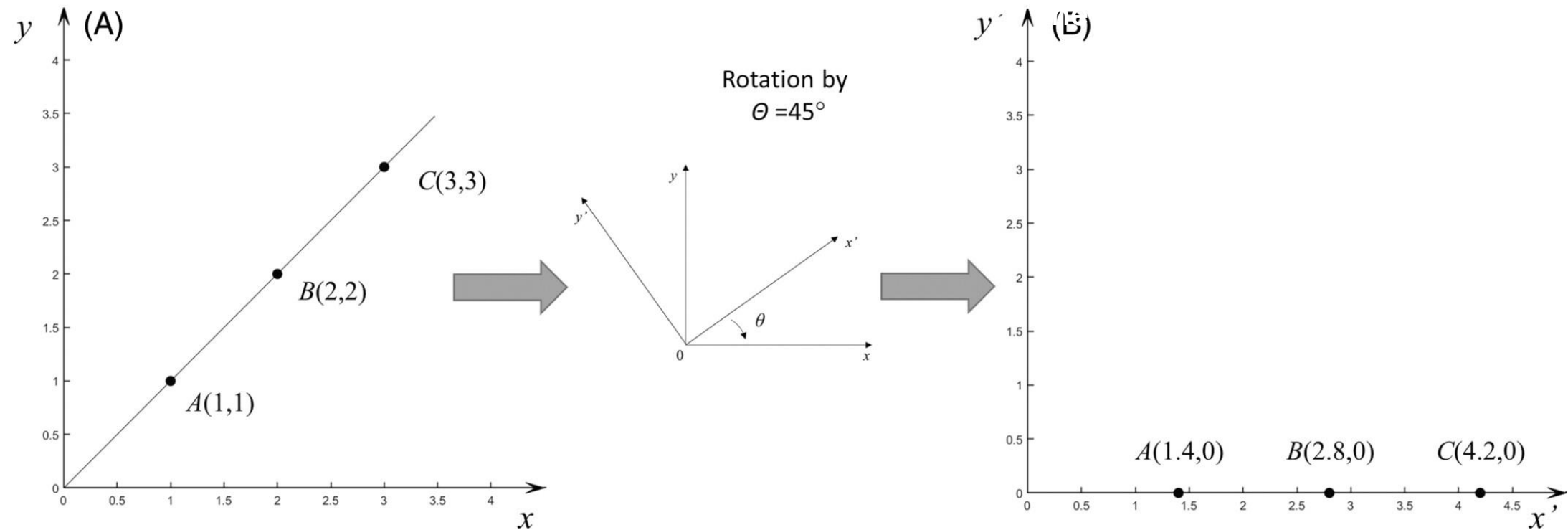
New coordinate system $x'Oy'$: by rotating the coordinates by $\theta=45^\circ$ counterclockwise

- $\text{var}(x') = 2$
- $\text{var}(y') = 0$

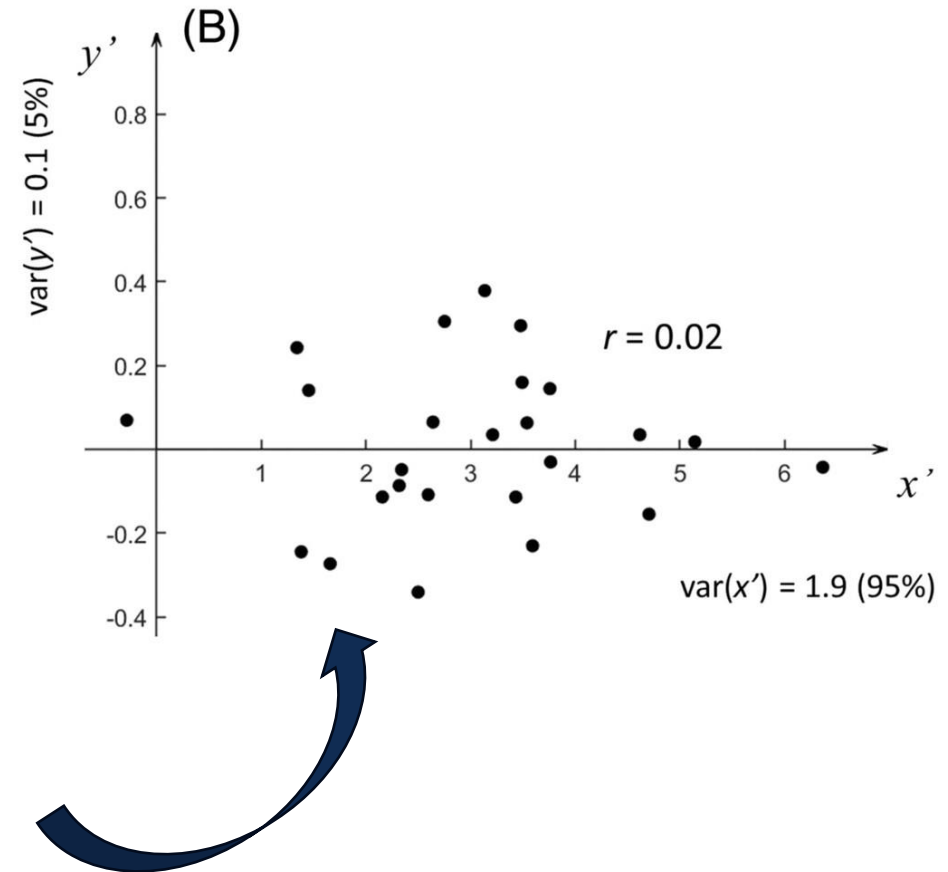
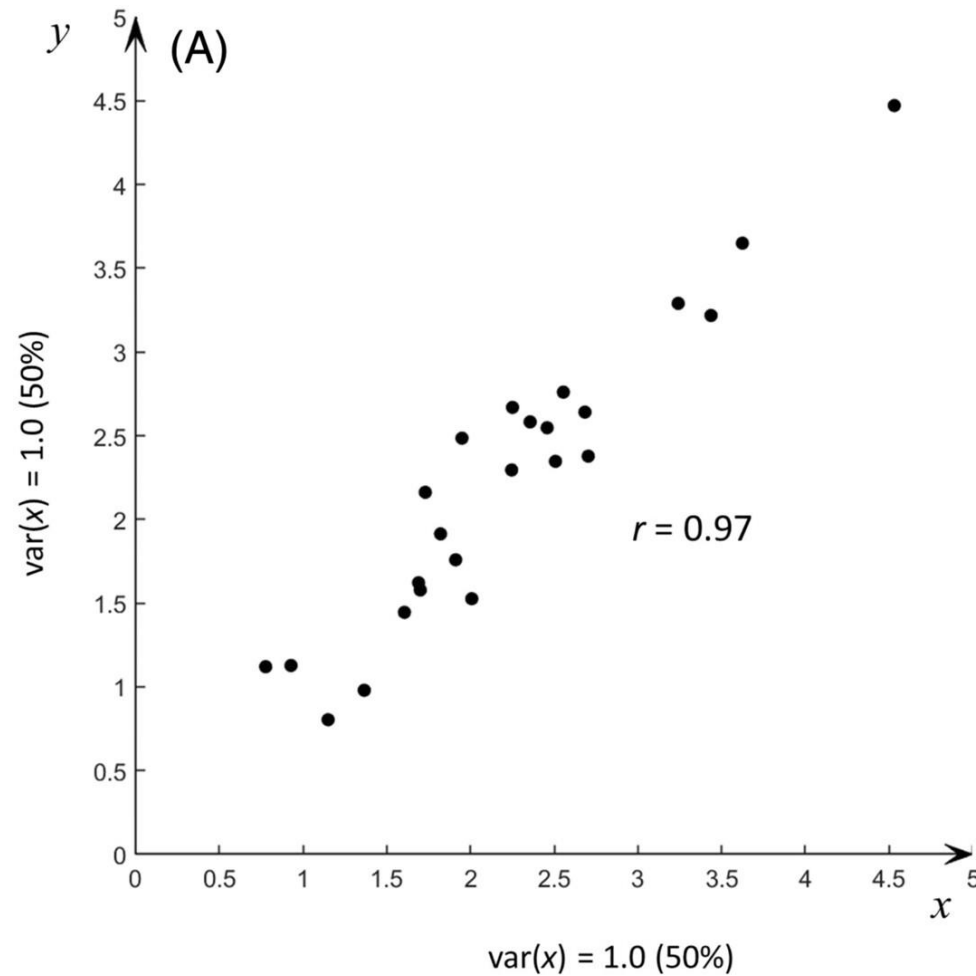
→ The variance is maximized along the x' direction.

PCA – From intuition to optimization

The **total variance is conserved**: $\text{var}(x) + \text{var}(y) = \text{var}(x') + \text{var}(y') = 2$.



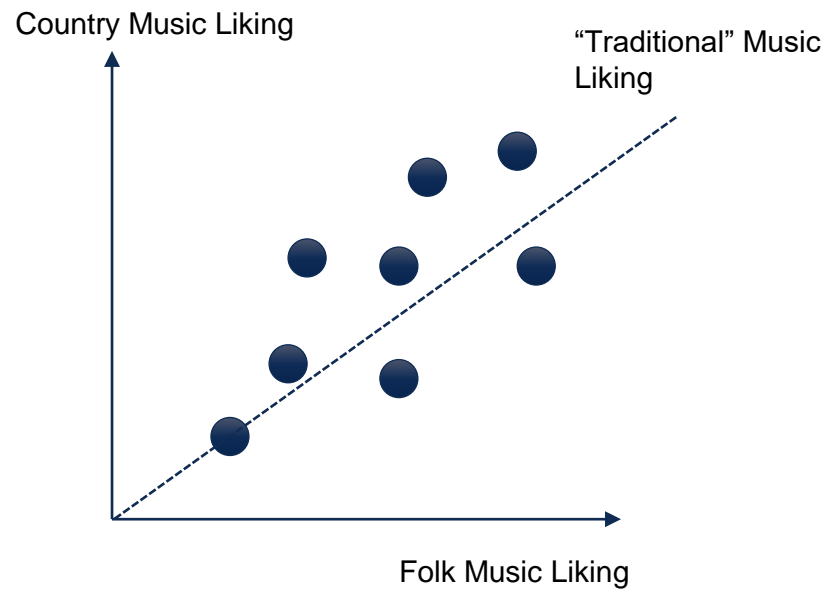
PCA – From intuition to optimization



PCA – From intuition to optimization

- **Principal Components Analysis (PCA)** starts extracting the maximum variance and puts them into the first component. After that, it removes that variance explained by the first component and then starts extracting maximum variance for the second component. This process goes to the last component.
- Technically, **PCA is an orthogonal linear transformation** that reorients the data to a novel coordinate system where the most significant variance, as determined by a scalar projection, aligns with the first coordinate (referred to as the first principal component). Subsequently, the second greatest variance aligns with the second coordinate, and this pattern continues for all the subsequent components.

PCA – From intuition to optimization

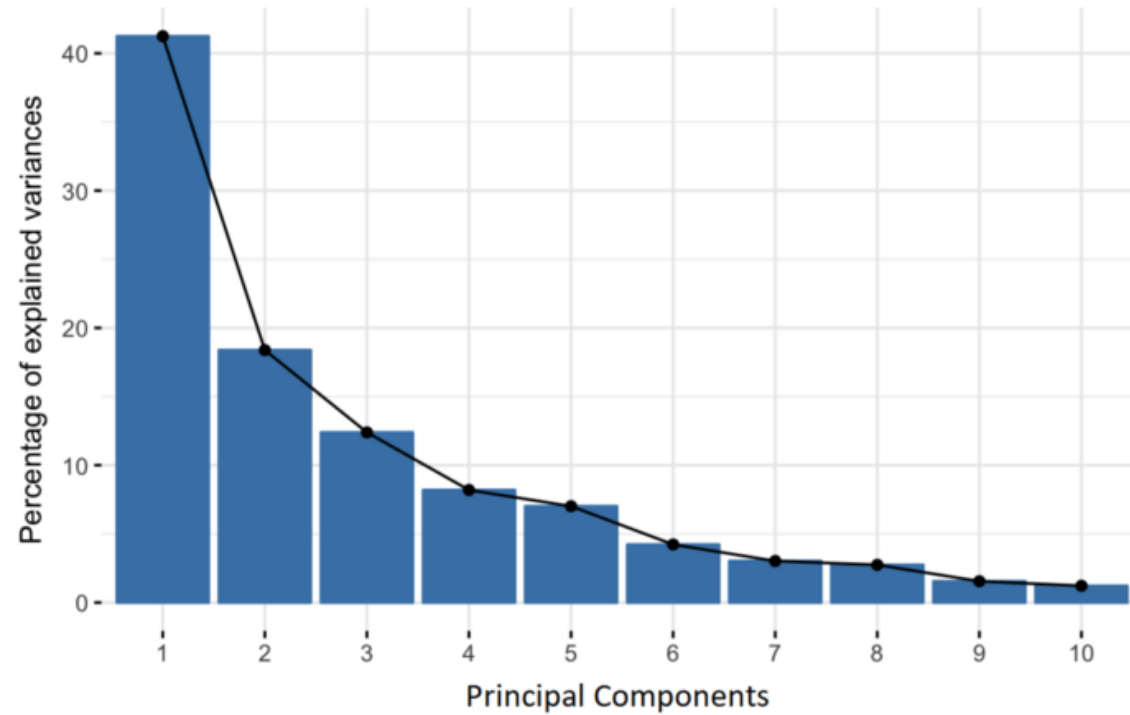


$$\begin{bmatrix} \text{Cov}(x,x) & \text{Cov}(x,y) \\ \text{Cov}(y,x) & \text{Cov}(y,y) \end{bmatrix}$$



The eigenvector associated with the largest eigenvalue indicates the direction in which the data has the most variance.

Principal Component Analysis



- New variables are uncorrelated
- Most of the information is contained in the first component

Dataset: Young People Survey

A **large-scale survey** was conducted **among young people** on many aspects of their lives, such as:

- Self perception
- Likes and fears
- Lifestyle
- Preferences in music, movie, etc.
- Some demographic information

In particular, the music preference was measured in a series of questions:

- I enjoy listening to music (Likert scale, 1 Not at all – 5 Very much)
- I prefer slow or fast songs (1 Slow – 5 Fast)
- I like the following music genres (1 Not at all – 5 Very much) – 17 genres

- | | | |
|----------------------|--------------------|------------------|
| • Dance, disco, funk | • Rock | • Rock n Roll |
| • Folk | • Metal, hard rock | • Alternative |
| • Country | • Punk | • Latin |
| • Classical | • Hip hop, rap | • Techno, Trance |
| • Musicals | • Reggae, Ska | • Opera |
| • Pop | • Swing, jazz | |

Likert Scales

- A Likert scale is a unidimensional **measurement tool** employed to gather the attitudes and opinions of respondents. This psychometric scale is frequently utilized to assess perspectives and sentiments regarding a brand, product, or target market.
- These scales facilitate the assessment of respondents' degrees of agreement or disagreement.
- The Likert scale assumes that the strength and intensity of experiences are linear. This implies a belief in the measurability of the variable along this continuum.

Strongly
disagree

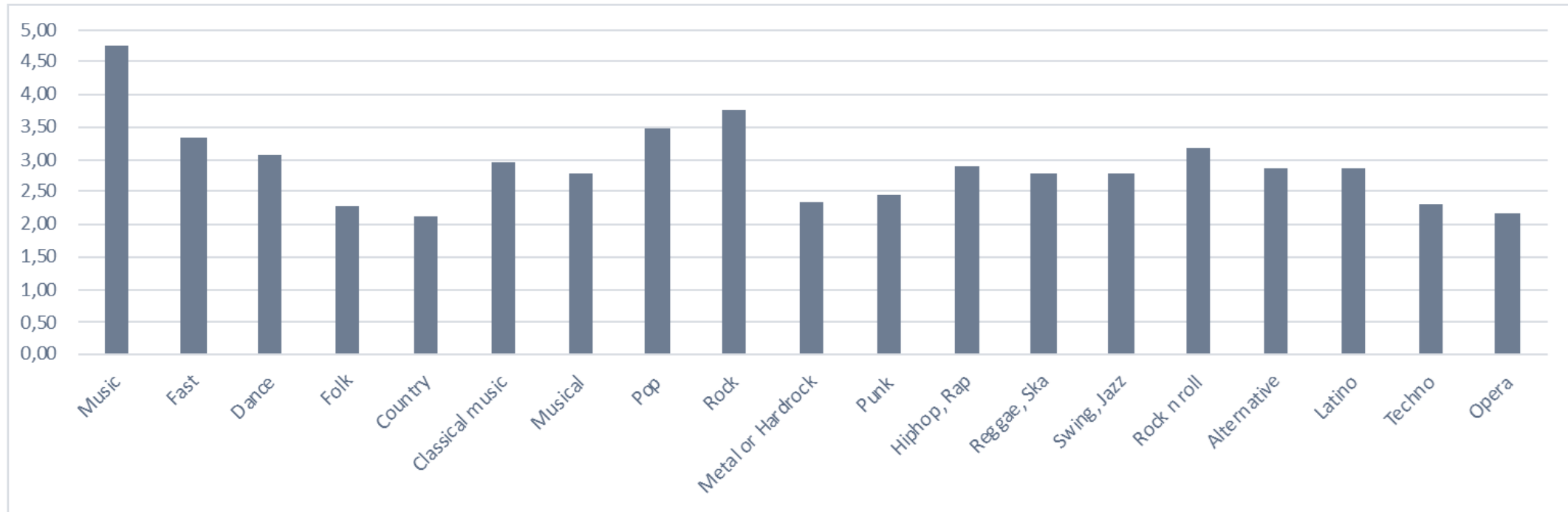
Disagree

Neutral

Agree

Strongly
agree

Dataset: Young People Survey



Dataset: Young People Survey

- Dance, disco, funk
- Folk
- Country
- Classical
- Musicals
- Pop
- Rock
- Metal, hard rock
- Punk
- Hip hop, rap
- Reggae, Ska
- Swing, jazz
- Rock'n Roll
- Alternative
- Latin
- Techno, Trance
- Opera

Summarizing the information in the
original variables with a (substantially)
lower number of components



- Component 1
- Component 2
- Component 3
- ...

PCA: Assumptions


Assumption #1: You have multiple variables that are measured at the continuous level (although ordinal data is very frequently used).

Typologies of data

- Categorical / Nominal
- Ordinal
- Continuous / Scale

Typologies of data

- Categorical / Nominal



Customer ID	Gender	Type of account	Free account	Basic account	Premium account	Nationality	Hair color
C1	Male	Free	Yes	No	No	Italian	Black
C2	Female	Basic	No	Yes	No	French	Brown
C3	Female	Basic	No	Yes	No	German	Blonde
C4	Male	Premium	No	No	Yes	Spanish	Brown
...


Customer ID	Gender	Type of account	Free account	Basic account	Premium account	Nationality	Hair color
C1	1	1	1	0	0	52	1
C2	0	2	0	1	0	23	2
C3	0	2	0	1	0	44	3
C4	1	3	0	0	1	79	2
...

Typologies of data

- Categorical / Nominal
 - The **numerical values** only represent distinctive groups and **do not carry any numerical meaning** (sum, average, etc. do not have meaning!)
 - The groups could be labeled with completely different numbers (but you need to **keep track of the labeling**)
 - Dummy variable (Free account, Basic account, and Premium account in the example) takes the value of 0 or 1 to indicate the absence or presence of the given category. To include variable with more than two categories in certain analysis, it needs to be transformed into dummy variables first.

Typologies of data

- Ordinal



Customer ID	Age	Education level	Household size	Driver's licence	Income level	BMI	Satisfaction level
C1	< 18	High school	> 5 persons	A	< 20K	Normal	Exceed exp.
C2	18-25	B.Sc.	1 person	B	20K – 30K	Underweight	Meets exp.
C3	26-35	Ph.D.	2 persons	D	30K – 40K	Overweight	Below exp.
C4	36-45	M.Sc.	3-5 persons	B	30K – 40K	Obese	Meets exp.
...

Customer ID	Age	Education level	Household size	Driver's licence	Income level	BMI	Satisfaction level
C1	1	1	4	1	1	2	3
C2	2	2	1	2	2	1	2
C3	3	4	2	4	3	3	1
C4	4	3	3	2	3	4	2
...

Typologies of data

- Ordinal
 - **There is certain order implied**, for example: young – old, lower education – higher education, less satisfied – more satisfied
 - The **difference in number**, however, does **not represent the meaning of distance between different levels** in the order
 - The numerical value could be arbitrarily assigned: you should keep track of the labeling

Typologies of data

- Continuous / Scale

Customer ID	Age	Height (cm)	BMI	Annual income (€)	Month since 1° purchase	Total spending (€)	Log-in per week
C1	16	168	20,1	5.000	3	50	27
C2	23	165	16,4	22.300	14	320	25
C3	31	173	26,7	38.000	22	760	12
C4	38	182	32,2	36.500	19	510	5
...

Typologies of data

- Continuous / Scale
 - There is an order implied
 - The **numbers and the difference between the numbers have actual measurable meanings**, such as: 2 years younger, 10cm taller, spend 300€ more.

PCA: Assumptions

Assumption #1: You have multiple variables that are measured at the continuous level (although ordinal data is very frequently used).

Assumption #2: There should be a linear relationship between all variables.

PCA: Assumptions

Assumption #1: You have multiple variables that are measured at the continuous level (although ordinal data is very frequently used).

Assumption #2: There should be a linear relationship between all variables.

Assumption #3: There should be no outliers.

PCA: Assumptions

Assumption #1: You have multiple variables that are measured at the continuous level (although ordinal data is very frequently used).

Assumption #2: There should be a linear relationship between all variables.

Assumption #3: There should be no outliers.

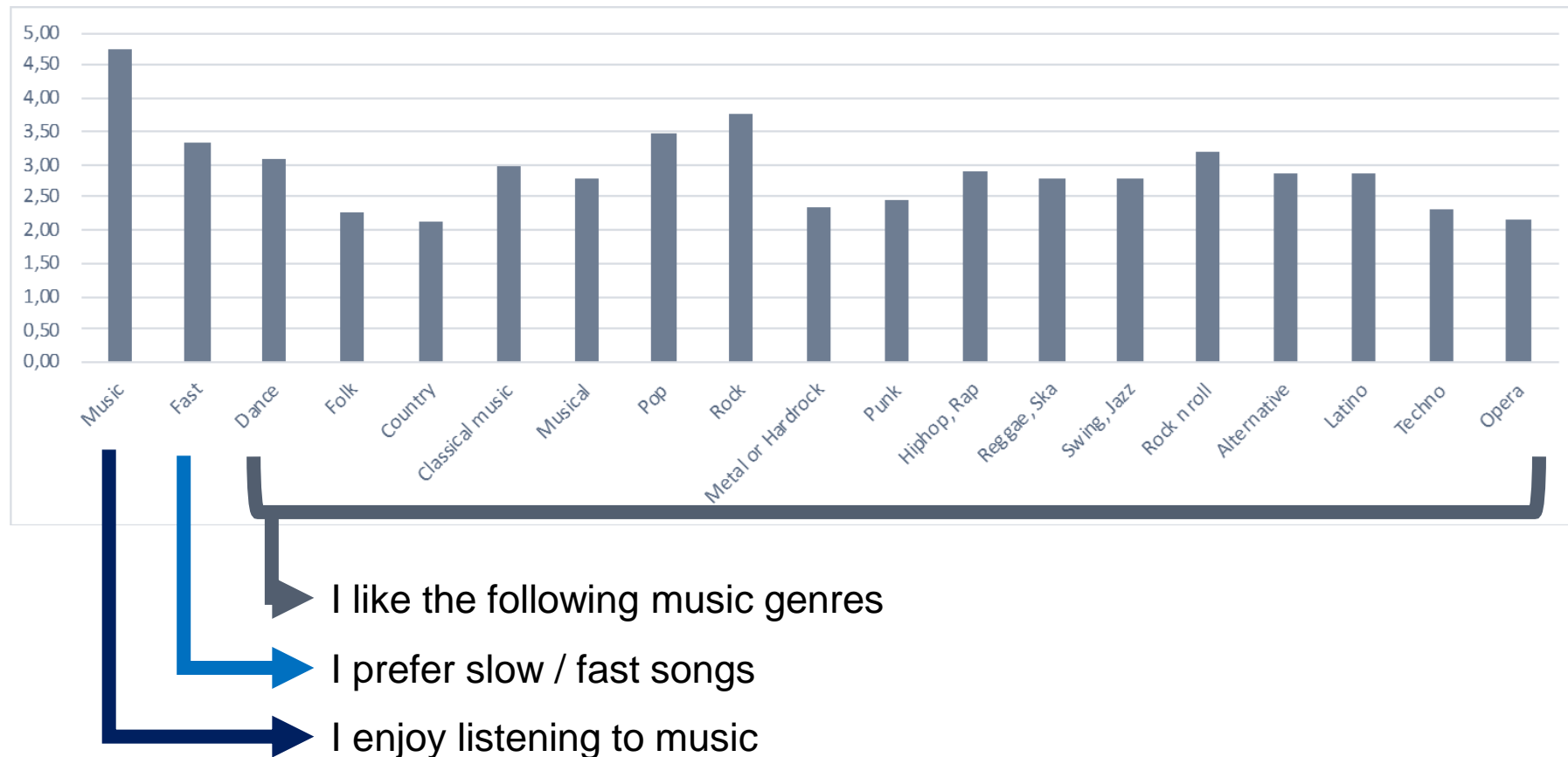
Assumption #4: There should be large sample sizes for a principal components analysis to produce a reliable result.

PCA: Steps

1. Variables selection
2. Rotation method identification
3. Number of principal component definition
4. Results interpretation

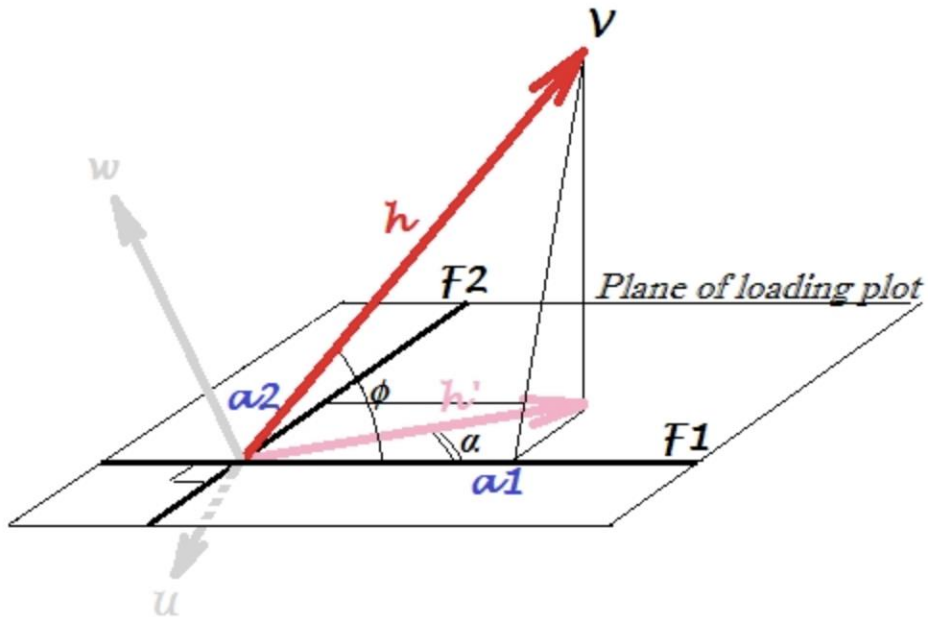
PCA: Variables Selection

Objective: to identify music preferences.



PCA: Rotation method identification

PCA loadings are the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed.



$$u_1 = \begin{bmatrix} 0.59 \\ 0.02 \\ 0.15 \end{bmatrix}$$

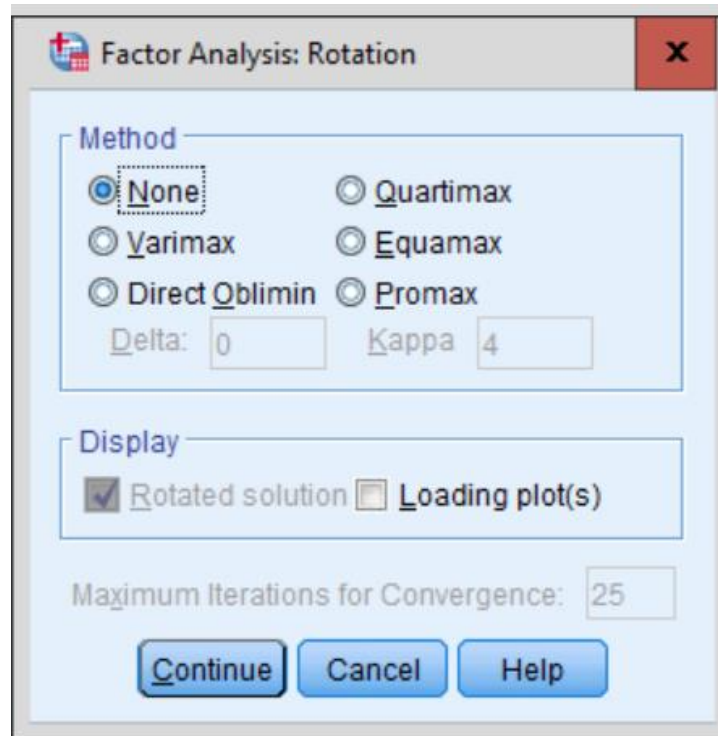
PCA: Rotation method identification

The goal of rotation is to improve the interpretability of the component solution by reaching simple structure.

In this case, loadings are the coefficients of the linear combination of the original variables from which the **rotated** principal components (PCs) are constructed.

	Componente				
	1	2	3	4	5
Dance	,006	-,115	,455	,681	,063
Folk	,671	-,001	,059	,043	,116
Country	,514	,167	,149	,071	,056
Classical music	,790	,109	-,083	-,068	,082
Musical	,518	,013	,517	-,203	,078
Pop	-,087	-,064	,793	,243	-,087
Rock	,091	,837	,106	-,133	,060
Metal or Hardrock	,124	,763	-,261	-,017	-,059
Punk	-,047	,775	-,133	-,001	,163
Hiphop, Rap	-,272	-,222	,224	,513	,373
Reggae, Ska	-,034	,154	,059	,142	,794
Swing, Jazz	,466	,126	,040	-,073	,637
Rock n roll	,268	,554	,187	-,149	,398
Alternative	,234	,411	-,327	-,054	,408
Latino	,297	-,142	,585	,045	,310
Techno	,051	-,021	-,067	,880	-,038
Opera	,788	,036	-,036	-,071	-,033

PCA: Rotation method identification



The goal of component rotation is to improve the interpretability of the factor solution by reaching simple structure

Orthogonal rotation (Varimax):

assumes that components are independent or uncorrelated with each other;

Oblique rotation (Oblimin):

assumes that components are not independent and are correlated

PCA: Number of principal component definition

As an unsupervised technique, the number of component is usually not known in advance; the «correct» number of components practically does not exist.

Though, some statistical indicators could help to define the number of components:

Factors with Eigen-value > 1 : the component explains more variances than a single variable

Scree plot: cut off at the «elbow» point, the value added by additional components is small

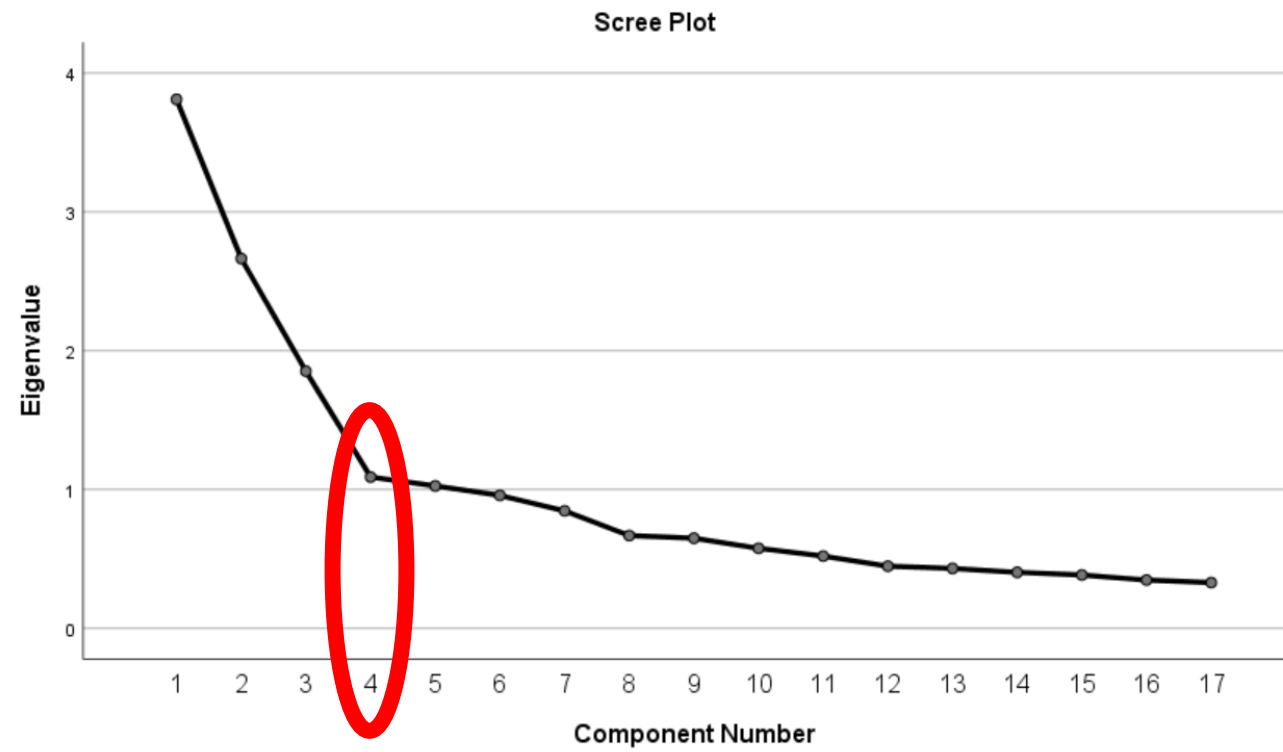
PCA: Number of principal component definition

Total Variance Explained						
Component	Total	Initial Eigenvalues		Extraction Sums of Squared Loadings		
		% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,811	22,415	22,415	3,811	22,415	22,415
2	2,663	15,667	38,082	2,663	15,667	38,082
3	1,852	10,894	48,976	1,852	10,894	48,976
4	1,000	6,110	55,285	1,000	6,110	55,285
5	1,025	6,032	61,418	1,025	6,032	61,418
6	,957	5,631	67,049			
7	,846	4,977	72,026			
8	,668	3,930	75,955			
9	,649	3,816	79,772			
10	,576	3,389	83,161			
11	,520	3,059	86,220			
12	,447	2,629	88,849			
13	,431	2,538	91,387			
14	,403	2,372	93,759			
15	,384	2,260	96,019			
16	,348	2,045	98,064			
17	,329	1,936	100,000			

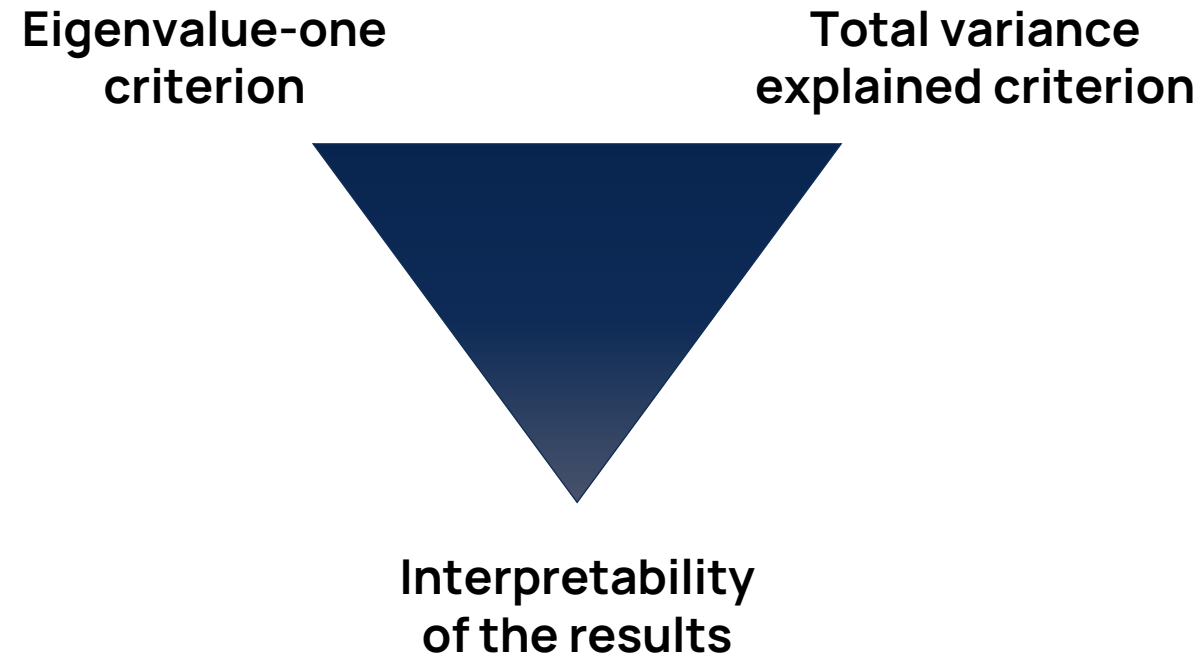
Extraction Method: Principal Component Analysis.

(Kaiser, 1960)

PCA: Scree plot



PCA: Number of principal component definition



PCA: Number of principal component definition

4-Factor solution

Rotated Component Matrix^a

	Component			
	1	2	3	4
Classicalmusic	,788			
Opera	,769			
Folk	,678			
SwingJazz	,562			
Country	,502			
Rock		,779		
Punk		,769		
Rocknroll		,668		
MetalorHardrock		,660		
Alternative		,553		
ReggaeSka		,478	,431	
Techno			,779	
HiphopRap			,666	
Dance			,665	
Pop				,738
Latino				,626
Musical	,512			,533

Extraction Method: Principal Component Analysis

5-Factor solution

Rotated Component Matrix^a

	Component				
	1	2	3	4	5
Classicalmusic	,790				
Opera	,788				
Folk	,671				
Musical	,518		,517		
Country	,514				
Rock		,837			
Punk		,775			
MetalorHardrock		,763			
Rocknroll		,554			
Alternative		,411			,408
Pop			,793		
Latino			,585		
Techno				,880	
Dance			,455	,681	
HiphopRap				,513	
ReggaeSka					,794
SwingJazz	,466				,637

Extraction Method: Principal Component Analysis

PCA: Results interpretation

5-Factor solution

Rotated Component Matrix^a

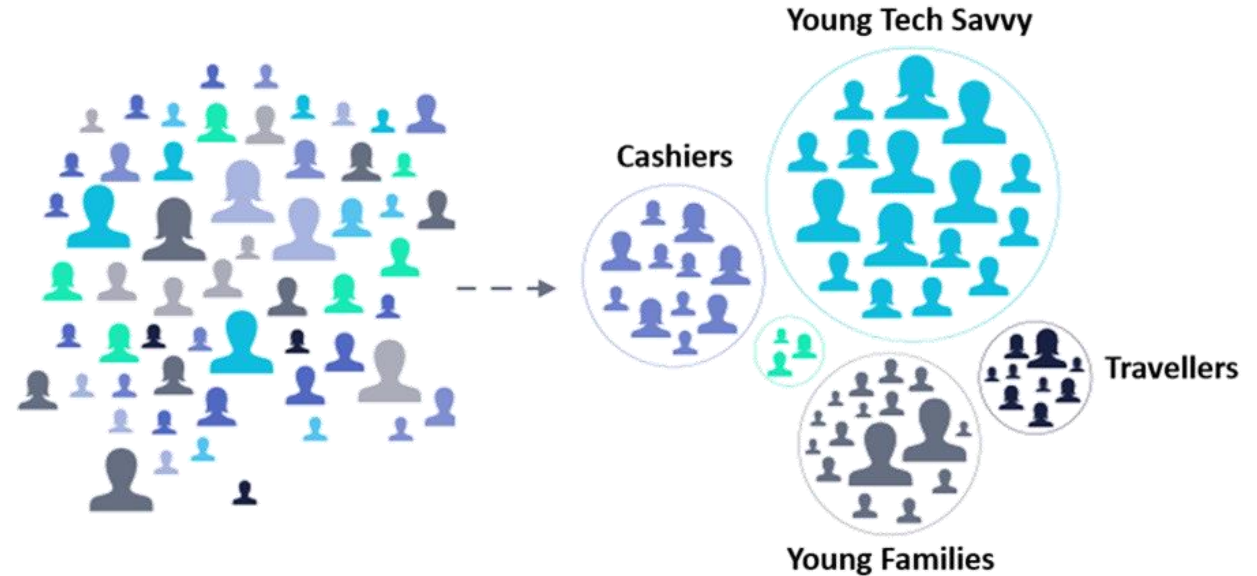
		Component				
		1	2	3	4	5
Classy	Classicalmusic	,790				
	Opera	,788				
	Folk	,671				
	Musical	,518		,517		
	Country	,514				
Rocky	Rock		,837			
	Punk		,775			
	MetalorHardrock		,763			
	Rocknroll		,554			
	Alternative		,411			,408
Dancy	Pop			,793		
	Latino			,585		
Disco	Techno				,880	
	Dance			,455	,681	
	HiphopRap				,513	
Jazzy	ReggaeSka					,794
	SwingJazz	,466				,637

Cluster Analysis

02

Cluster Analysis

Subdivide the records of a dataset into **homogeneous groups of observations**, called clusters, so that observations belonging to one group are similar to one another and dissimilar from observations included in other groups.



Taxonomy of clustering methods

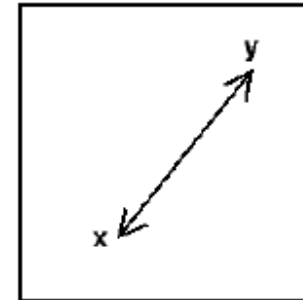
Clustering methods can be classified into a few main types based on the **logic used for deriving the clusters**:

- Partition methods
- Hierarchical methods
- Density based methods
- Grid methods

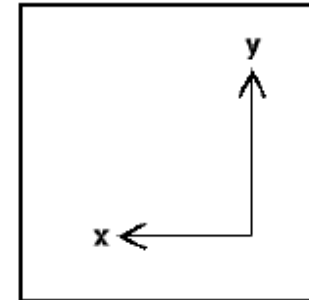
Partition methods

Partitioning methods divide the data into a **predefined number of clusters** by optimizing a given criterion (such as minimizing the distance between data points within the same cluster). They are characterized by:

- Fixed number of clusters
- Flat structure
- Iterative optimization



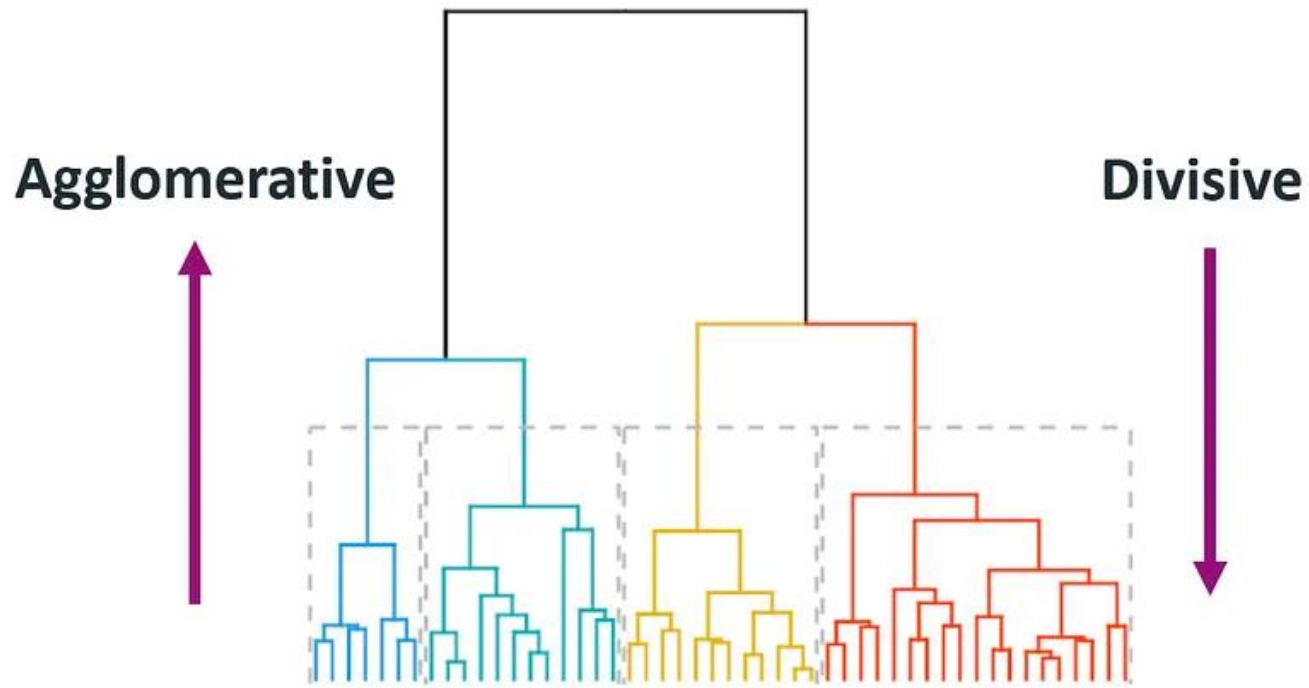
Euclidean



Manhattan

Hierarchical methods

Hierarchical clustering builds a hierarchy or tree-like structure of clusters (represented in a **dendrogram**) showing how clusters are formed at different levels of granularity. These methods are characterized by:

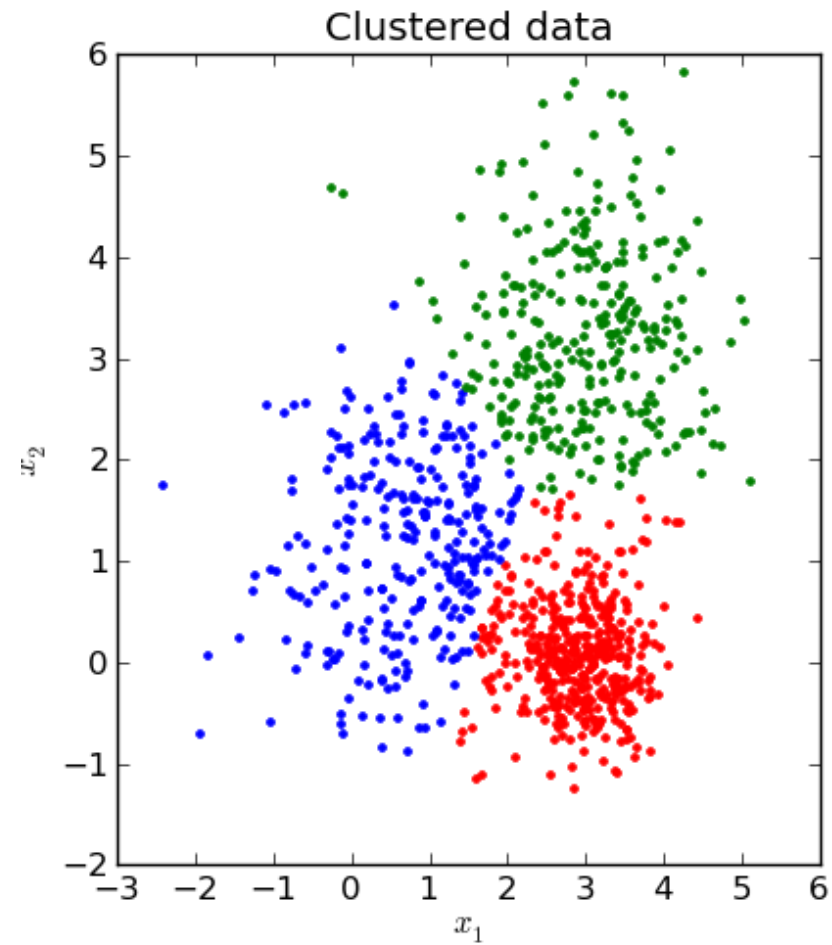
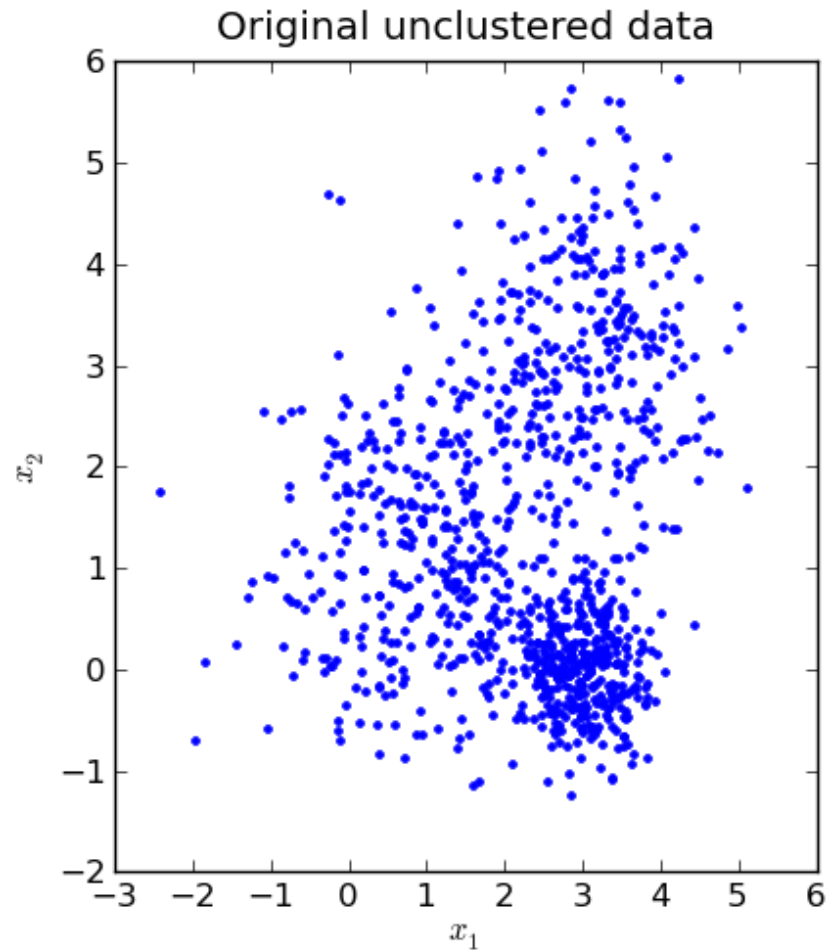


- No need to specify the number of clusters upfront
- Hierarchical structure
- Agglomerative vs. divisive approach

Clustering methods: K-Means

- Clustering objects into homogeneous groups based on variables with **continuous/scale variables** (e.g. age, height).
- It classifies objects by computing the **distance** among the objects. Variables with very different scales should be standardized to avoid the clustering being dominated by the measurement scale of a large range.
- The **number of clusters** has to be imposed before computation: several K-Means cluster analyses could be performed to choose the best solution
- It is fast and easy to perform on a large dataset
- It cannot classify subjects with missing values

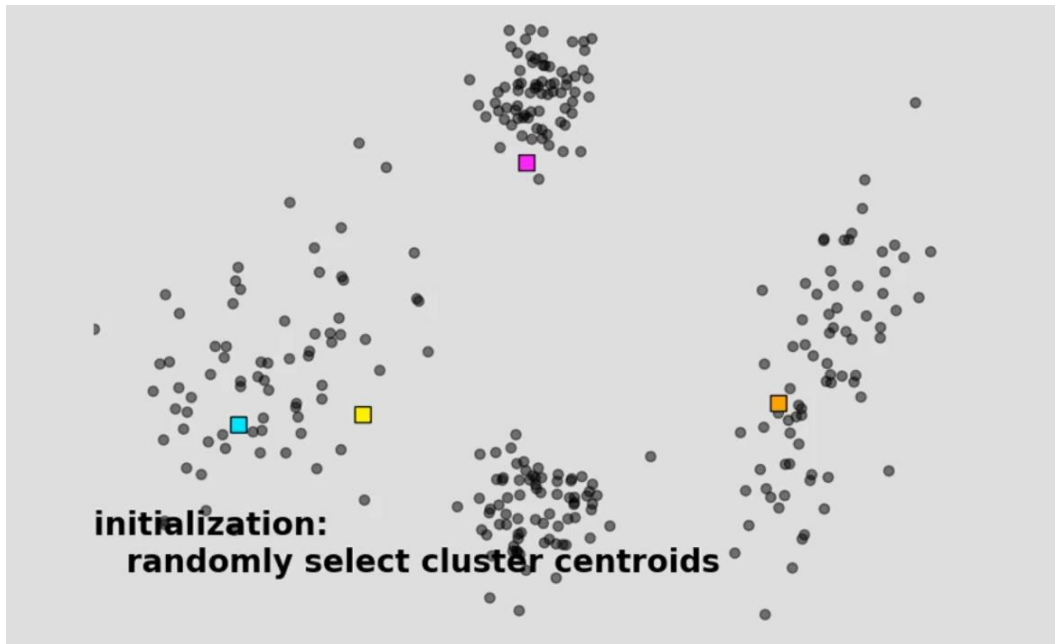
Clustering methods: K-Means



K-Means clustering: Steps

1. Variables selection
2. Number of clusters identification
3. Convergence assessment
4. Robustness assessment
5. Results interpretation

Clustering methods: K-Means



Step 1: Choose the number of clusters k

Step 2: Select k random centroids

Step 3: Assign all the points to the closest cluster centroid

Step 4: Recompute the centroids of newly formed clusters

Stop If:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

Dataset: Young People Survey

A **large-scale survey** was conducted **among young people** on many aspects of their lives, such as:

- Self perception
- Likes and fears
- Lifestyle
- Preferences in music, movie, etc.
- Some demographic information

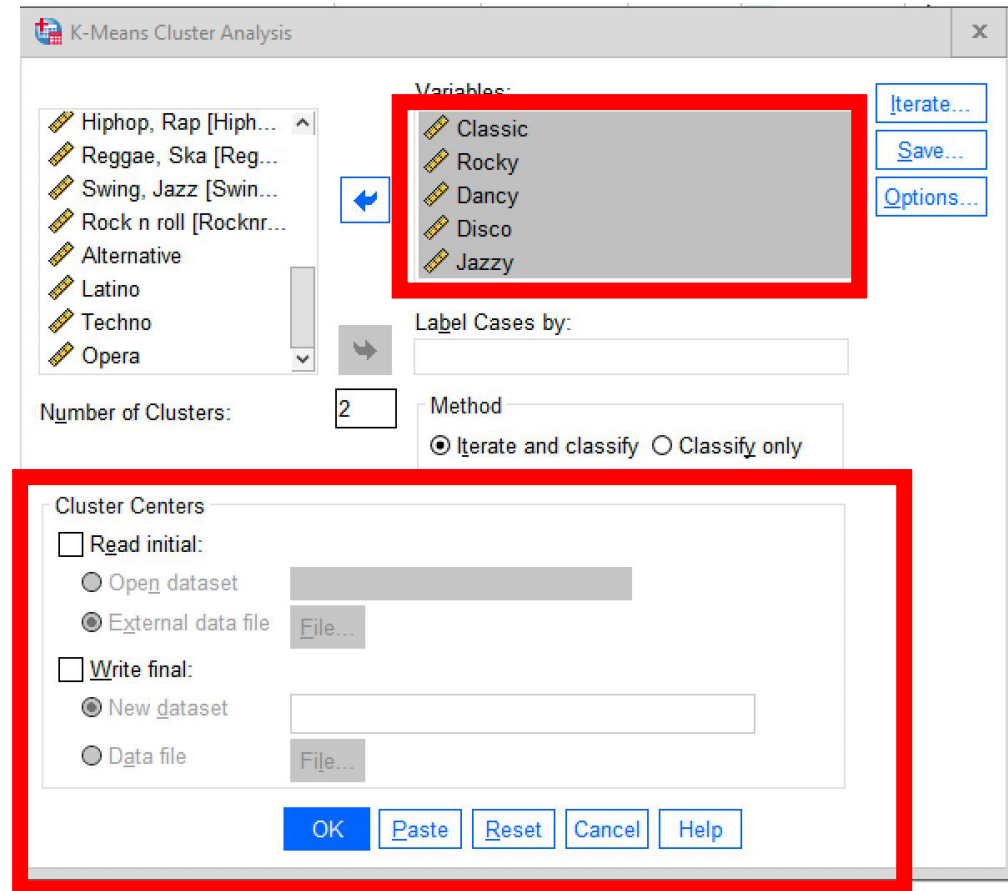
In particular, the music preference was measured in a series of questions:

- I enjoy listening to music (Likert scale, 1 Not at all – 5 Very much)
- I prefer slow or fast songs (1 Slow – 5 Fast)
- I like the following music genres (1 Not at all – 5 Very much) – 17 genres

- | | | |
|----------------------|--------------------|------------------|
| • Dance, disco, funk | • Rock | • Rock n Roll |
| • Folk | • Metal, hard rock | • Alternative |
| • Country | • Punk | • Latin |
| • Classical | • Hip hop, rap | • Techno, Trance |
| • Musicals | • Reggae, Ska | • Opera |
| • Pop | • Swing, jazz | |

K-Means clustering: Variables selection

Not all the variables in a dataset should be put together in a cluster analysis. You should consider the relationships between the input variables



K-Means: Defining k

Convergence

Robustness



**Interpretability
of the results**

K-Means clustering: Convergence assessment

Iteration History^a

Change in Cluster Centers		
Iteration	1	2
1	2,994	3,275
2	,087	,126
3	,074	,095
4	,059	,068
5	,035	,039
6	,035	,039
7	,040	,045
8	,037	,042
9	,047	,053
10	,062	,071
11	,077	,093
12	,073	,087
13	,063	,077
14	,040	,048
15	,039	,045
16	,043	,049
17	,026	,030
18	,031	,035
19	,033	,036
20	,034	,039

a. Iterations stopped because

Iteration History^a

Change in Cluster Centers					
Iteration	1	2	3	4	5
1	1,932	2,589	1,944	2,057	2,361
2	,192	,404	,157	,177	,461
3	,153	,225	,113	,200	,289
4	,119	,165	,096	,127	,152
5	,096	,115	,116	,178	,121
6	,053	,092	,117	,139	,055
7	,067	,097	,082	,080	,037
8	,064	,079	,071	,097	,065
9	,054	,082	,115	,151	,075
10	,026	,051	,099	,130	,149
11	,051	,046	,064	,050	,068
12	,021	,027	,023	,040	,055
13	,022	,022	,022	,022	,022
14	,000	,000	,000	,000	,000

K-Means clustering: Robustness assessment

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Classy	87,237	4	,593	848	147,056	,000
Rocky	59,897	4	,722	848	82,940	,000
Dancy	71,404	4	,668	848	106,907	,000
Disco	120,142	4	,438	848	274,292	,000
Jazzy	72,332	4	,664	848	109,011	,000

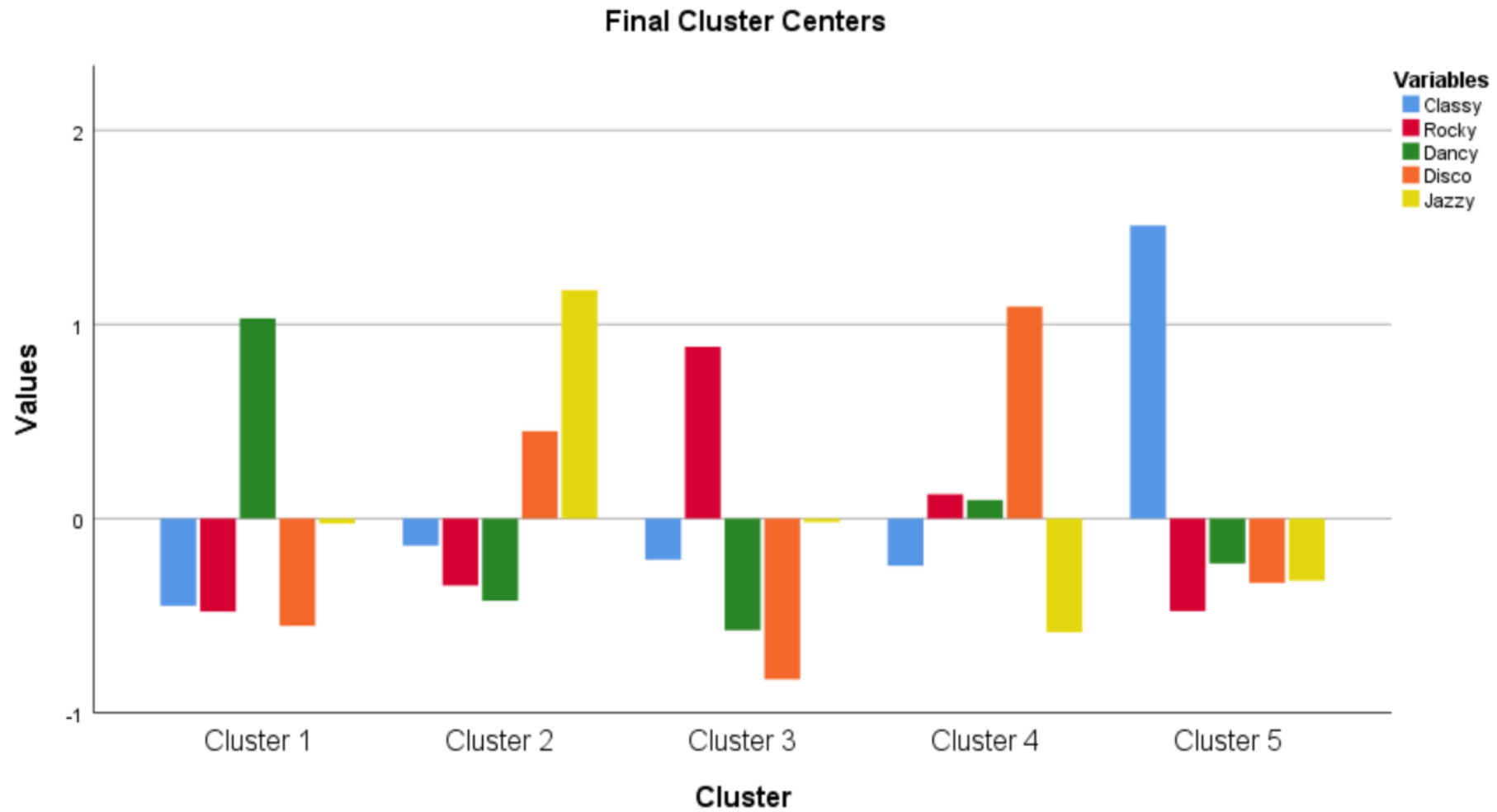
The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

K-Means clustering: Robustness & Interpretability of the Results

**Number of Cases in
each Cluster**

Cluster	1	176,000
	2	147,000
	3	191,000
	4	212,000
	5	127,000
Valid		853,000

K-Means clustering: Interpretability of the Results



K-Means clustering: Results interpretation

Which solution to choose?

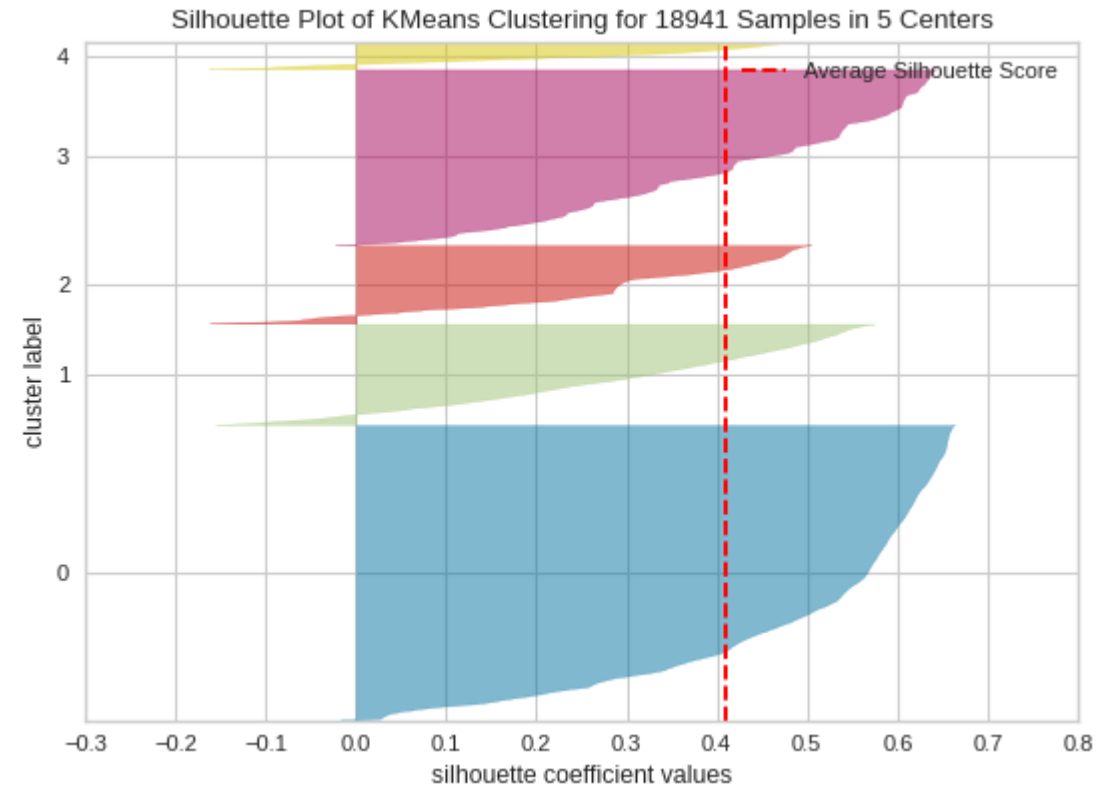
- **Statistically**, the best solution can be suggested by indicator of distinctiveness among the clusters

Silhouette coefficient

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$a(i)$: The average distance from the i th point to the other points in the same cluster.

$b(i)$: The minimum average distance from the i th point to points in a different cluster, minimized over clusters.



K-Means clustering: Results interpretation

Which solution to choose?

- **Statistically**, the best solution can be suggested by indicator of distinctiveness among the clusters
- For **marketing application**, it is more important to consider:
 - Whether the clusters provide you **relevant and actionable insights**
 - Whether these **insights are consistent with your business**
 - Whether it satisfies the criteria of '**good segmentation**'

K-Means clustering: Results interpretation

From experience, there could be some general consideration when it comes to selecting the right clustering solution:

- Solution with **too few clusters** may not adequately capture the diversity (for example, the 2-cluster solution of music preference only found substantial difference regarding classical music, while the preferences for all other types remain average)
- Solution with **too many clusters** will be increasingly difficult to interpret and to manage implementation of eventual applications



POLITECNICO
MILANO 1863

Gloria Peggiani

AY 2024/2025