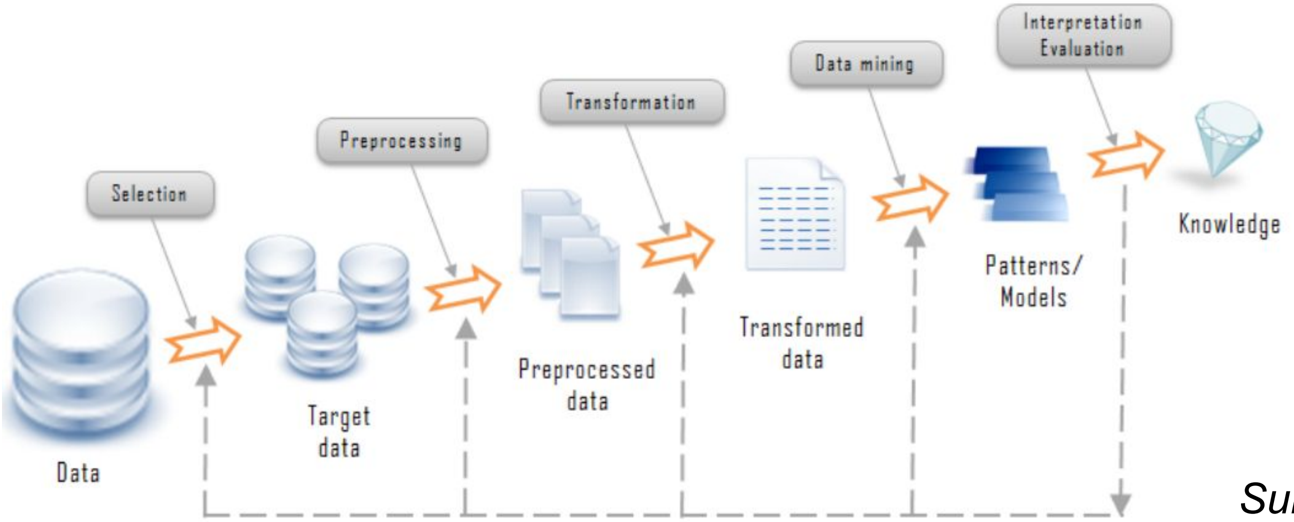
EI5IS102 Traitement de l'Information

Lecture 1:
Principal Component Analysis

Charles Brazier
Postdoctoral researcher
Université de Bordeaux, CNRS, Bordeaux INP, LaBRI
France

The data analysis process

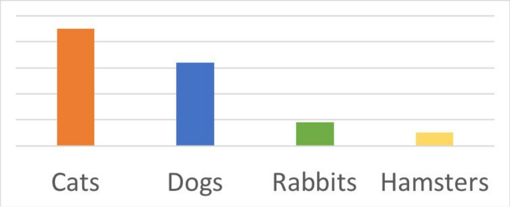


Sumiran, 2018

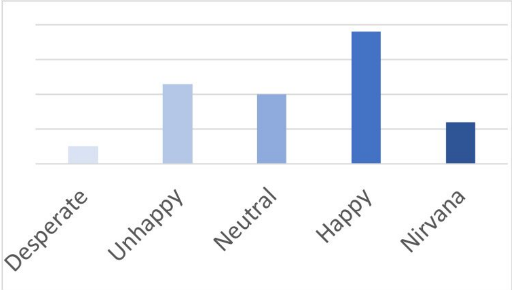
Types of data

Qualitative (categories)

Nominal:

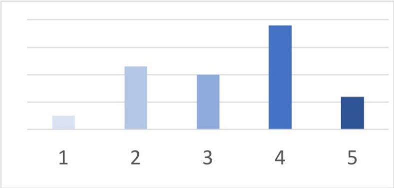


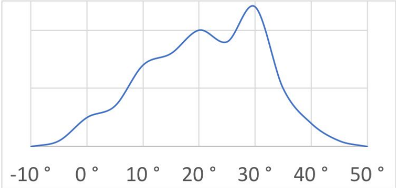
Ordered:



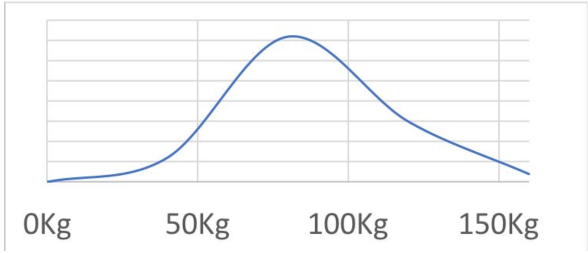
Quantitative (numerical values)

Interval (discrete or continuous):





Ratio:



Content of this course

How to represent data represented by high-dimensional data tables ?

Question 1: which data?
- Quantitative data: numerical values
- Qualitative data: categories, textual data, questionnaires, etc.

Question 2: which analysis?
- Quantitative data: PCA (Principal Component Analysis)
- Qualitative data: MCA (Multiple Correspondence Analysis)

Question 3: how to visualize the data?
- Dimension reduction: preprocessing data for clustering and classification

Content of this course

How to represent data represented by high-dimensional data tables ?

Question 1: which data?
- Quantitative data: numerical values
- Qualitative data: categories, textual data, questionnaires, etc.

Question 2: which analysis?
- Quantitative data: **PCA** (Principal Component Analysis)
- Qualitative data: MCA (Multiple Correspondence Analysis)

Question 3: how to visualize the data?
- Dimension reduction: preprocessing data for clustering and classification

Multidimensional quantitative data

Multidimensional data: data with multiple attributes, each represented as a dimension

Example 1: Composition of different water brands

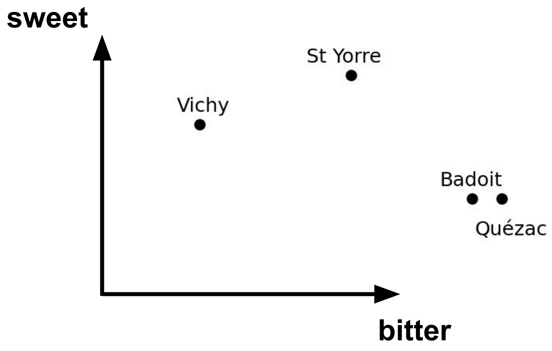
Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

Multidimensional quantitative data

Multidimensional data: data with multiple attributes, each represented as a dimension

Example 1: Composition of different water brands

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

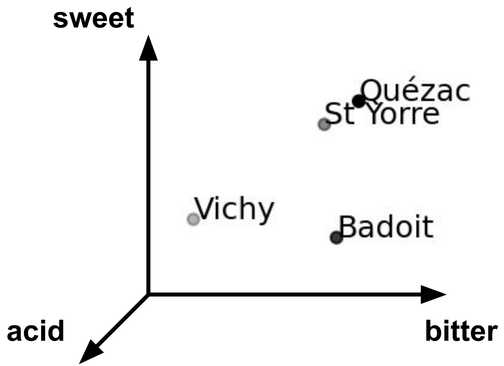


Multidimensional quantitative data

Multidimensional data: data with multiple attributes, each represented as a dimension

Example 1: Composition of different water brands

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

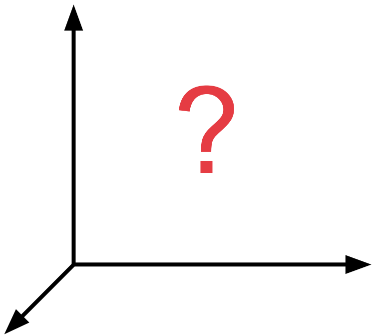


Multidimensional quantitative data

Multidimensional data: data with multiple attributes, each represented as a dimension

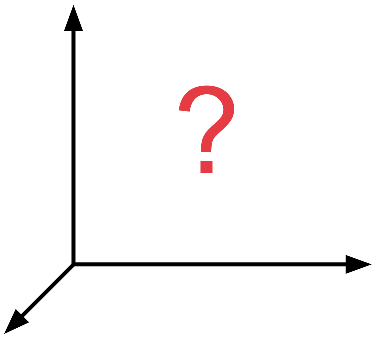
<u>Example 1:</u> Composition of different water brands

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3



Multidimensional quantitative data

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3



Objectives:

- How to visualize and interpret large datasets
- How to reduce high-dimensional data while preserving important information

Answered questions:

- What brands can be considered similar?
- Which attributes are discriminating or redundant?
- …

Multidimensional quantitative data

variables

instances

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

What will be computed during an analysis?

- Distance matrix between **observations/instances**

	St Yorre	Badoit	Vichy	Quézac
St Yorre	0.000000	1.852026	1.063015	2.109502
Badoit	1.852026	0.000000	1.788854	1.122497
Vichy	1.063015	1.788854	0.000000	2.481935
Quézac	2.109502	1.122497	2.481935	0.000000

- Correlation matrix between **attributes/variables**

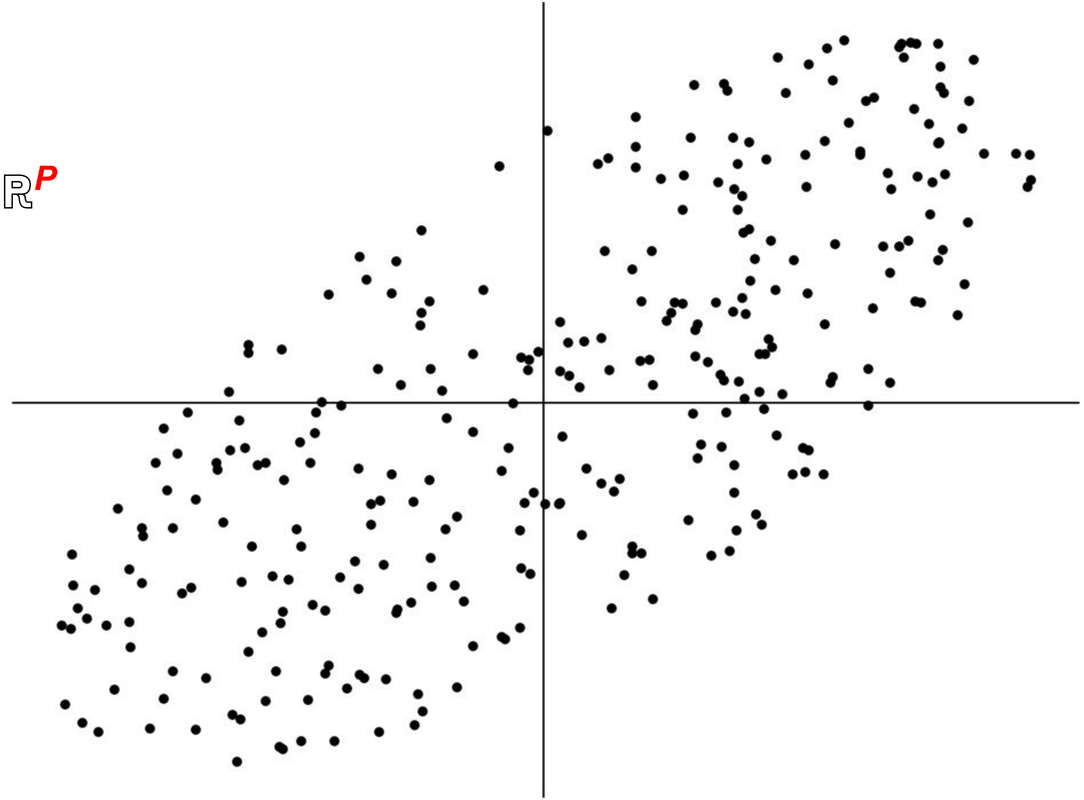
	bitter	sweet	acid	salted	alcaline
bitter	1.000000	-0.688487	0.812214	-0.790405	-0.966917
sweet	-0.688487	1.000000	-0.425165	0.988248	0.787839
acid	0.812214	-0.425165	1.000000	-0.518347	-0.860243
salted	-0.790405	0.988248	-0.518347	1.000000	0.863737
alcaline	-0.966917	0.787839	-0.860243	0.863737	1.000000

- …

Principal Component Analysis

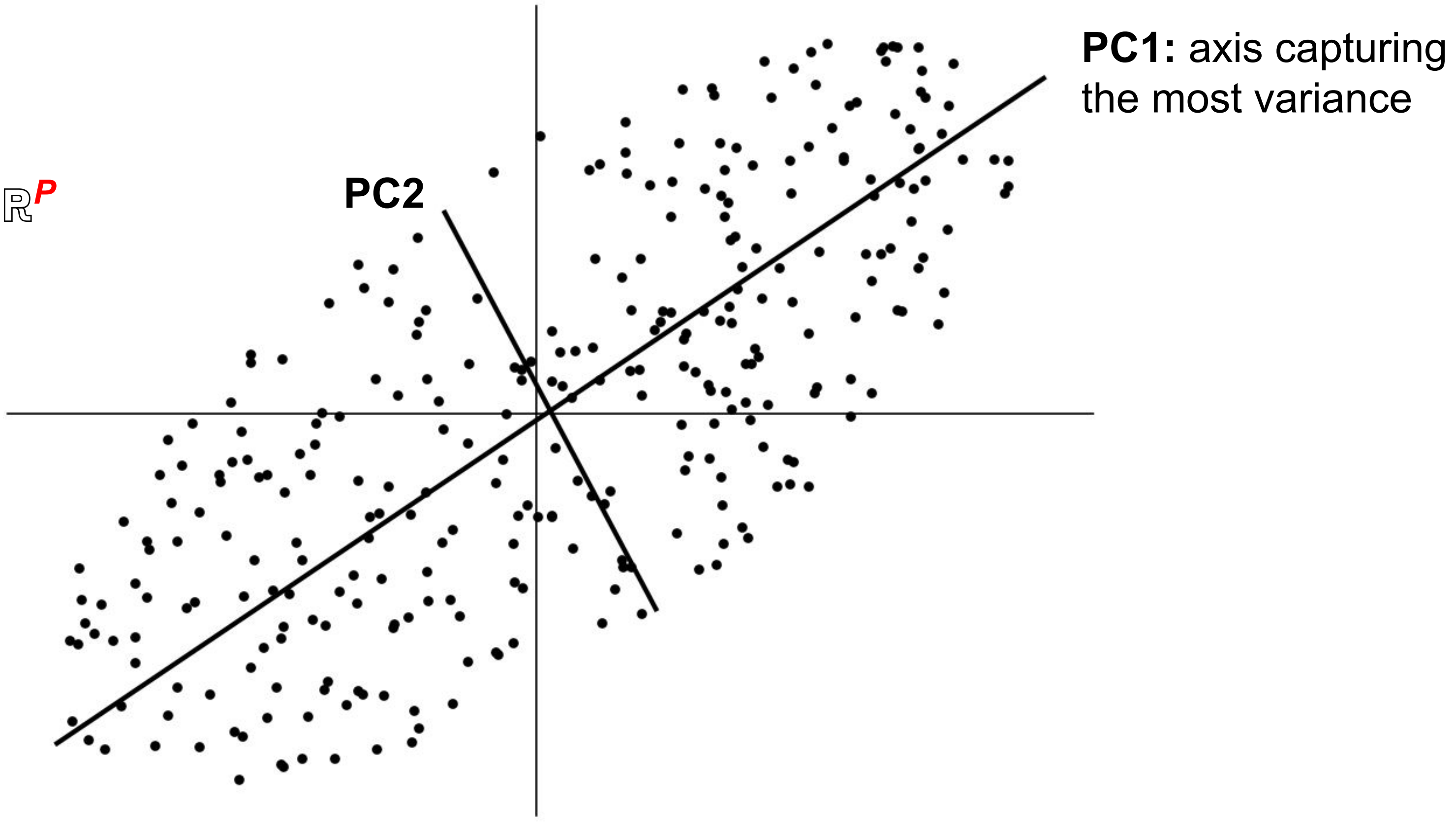
Principal Component Analysis

Data points in $\mathbb{R}^{P}$



Principal Component Analysis

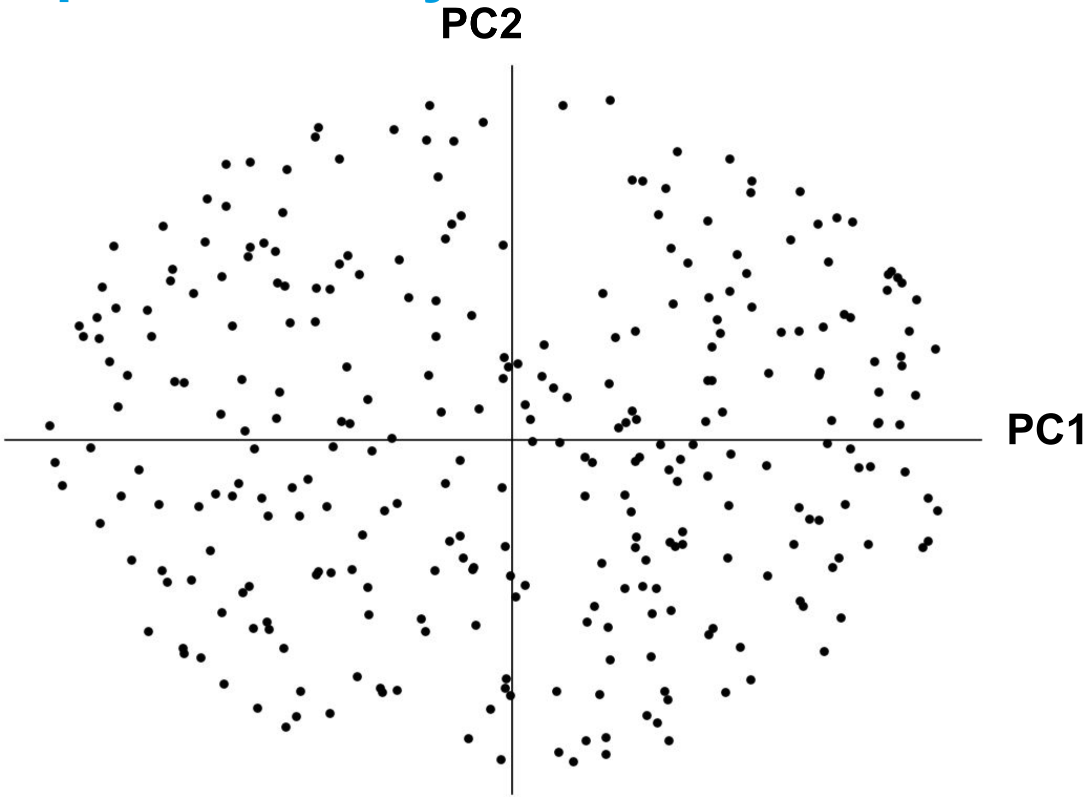
Data points in $\mathbb{R}^P$



Principal Components

Principal Component Analysis

Data points in $\mathbb{R}^2$

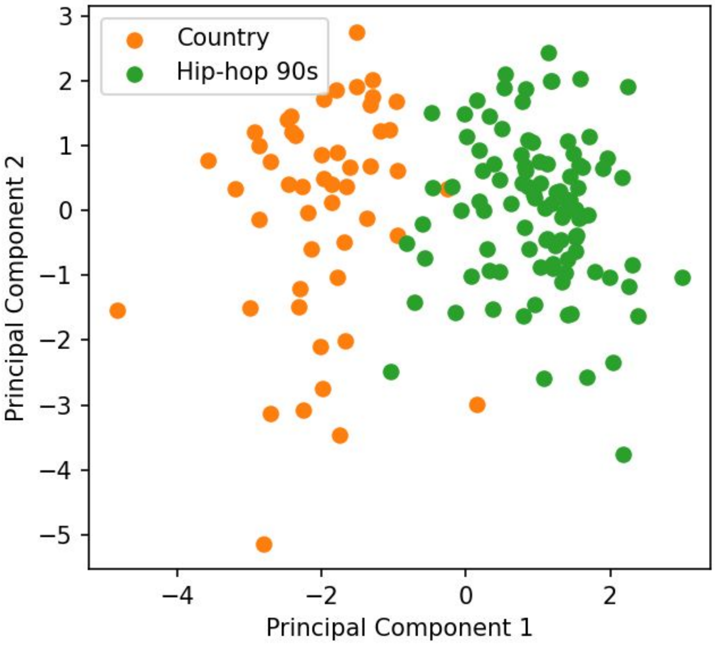


Dimension reduction

Principal Component Analysis

Why PCA?

- Data visualization: from many to 2D plots
- Noise reduction: identify and remove less important dimensions
- Help for classification and clustering



Notations

Objective: visualize the data cloud of **instances** described by different **variables**

- x_i^j : observation of **variable j** of **instance i**
- n: number of instances
- p: number of variables

Multidimensional quantitative data represented by a matrix of n lines and p columns

$$X = \begin{pmatrix} x_1^1 & \cdots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^p \end{pmatrix}$$

n=4

p=5

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

⚠️ n and p can be very large!

Data analysis with music data

name	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	sign.
190306-1...	Hideyuki...	0.43100	0.02030	9	-27.7570	0.048400	0.993000	0.940000	0.09060	0.26600	175.4410	155533	3
1st of T...	Bone Thu...	0.72900	0.58100	5	-8.2350	0.180000	0.077700	0.000004	0.69600	0.50800	74.0380	314680	4
28	Zach Bry...	0.49200	0.51900	7	-6.8860	0.028800	0.227000	0.000085	0.06930	0.43500	80.8680	233333	3
93 'Til ...	Souls Of...	0.59000	0.67200	1	-11.7920	0.412000	0.125000	0.000001	0.14700	0.68800	206.2470	286440	4
A Bar So...	Shabooze...	0.72200	0.70900	9	-4.9500	0.027300	0.063300	0.000000	0.08040	0.60400	81.0120	171292	4
A Cigare...	Gavin Ad...	0.61000	0.15000	7	-11.9460	0.028400	0.855000	0.000002	0.12800	0.21000	136.0890	179883	4
A New St...	Ferragno	0.35000	0.01500	8	-28.4700	0.050900	0.994000	0.958000	0.11500	0.29300	124.7840	155040	4
A Safe S...	Aramis M...	0.42200	0.01560	10	-28.8430	0.044900	0.993000	0.919000	0.10600	0.24800	115.4160	148299	4
A quiet ...	Christia...	0.45700	0.00836	7	-28.7330	0.039500	0.995000	0.955000	0.08570	0.28500	54.6210	122445	4
A tale t...	Luiza Sc...	0.38600	0.01280	1	-28.6600	0.039600	0.991000	0.931000	0.10900	0.18900	126.3740	130134	4
ATLiens	Outkast	0.91800	0.73400	11	-2.8320	0.269000	0.029600	0.000008	0.19100	0.60800	97.0440	230693	4
Adieux	Ludovico...	0.22500	0.00407	3	-37.1070	0.046400	0.987000	0.936000	0.11200	0.32000	73.5950	175493	4
Afterlig...	Arlo Thi...	0.38300	0.01430	1	-30.2790	0.044400	0.993000	0.914000	0.09750	0.18300	71.0290	200250	3
Ain't No...	Luke Com...	0.48700	0.65800	5	-9.9730	0.029600	0.008770	0.007770	0.10000	0.28200	142.1800	210950	4
Almenno ...	Rhian Ca...	0.43400	0.01200	2	-35.8260	0.032100	0.992000	0.896000	0.10600	0.29300	107.4370	167189	4
Almost A...	Florenti...	0.43000	0.06880	7	-27.3940	0.026800	0.990000	0.909000	0.10700	0.21400	83.5730	173133	4
Am I Oka...	Megan Mo...	0.59300	0.73400	9	-5.4160	0.051400	0.020000	0.000000	0.13800	0.51800	125.9370	235003	4
Ante Up ...	M.O.P.	0.69900	0.79300	1	-4.8560	0.268000	0.005140	0.000003	0.70000	0.92900	94.1380	248693	4
Arabesco	Lorenzo ...	0.36800	0.00394	3	-34.8920	0.050600	0.991000	0.903000	0.11400	0.13900	103.9680	155944	4
Arbor	Samuel K...	0.35200	0.00519	2	-34.3590	0.043600	0.995000	0.963000	0.09810	0.41400	89.4750	155500	4

...

Spotify API: 250 instances (n=250), 12 variables (p=12)

Data analysis with music data

$\mathbf{x}^1 \in \mathbb{R}^n$

$\mathbf{x}_3 \in \mathbb{R}^p$

name	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	sign.
190306-1...	Hideyuki...	0.43100	0.02030	9	-27.7570	0.048400	0.993000	0.940000	0.09060	0.26600	175.4410	155533	3
1st of T...	Bone Thu...	0.72900	0.58100	5	-8.2350	0.180000	0.077700	0.000004	0.69600	0.50800	74.0380	314680	4
28	Zach Bry...	0.49200	0.51900	7	-6.8860	0.028800	0.227000	0.000085	0.06930	0.43500	80.8680	233333	3
93 'Til ...	Souls Of...	0.59000	0.67200	1	-11.7920	0.412000	0.125000	0.000001	0.14700	0.68800	206.2470	286440	4
A Bar So...	Shabooze...	0.72200	0.70900	9	-4.9500	0.027300	0.063300	0.000000	0.08040	0.60400	81.0120	171292	4
A Cigare...	Gavin Ad...	0.61000	0.15000	7	-11.9460	0.028400	0.855000	0.000002	0.12800	0.21000	136.0890	179883	4
A New St...	Ferragno	0.35000	0.01500	8	-28.4700	0.050900	0.994000	0.958000	0.11500	0.29300	124.7840	155040	4
A Safe S...	Aramis M...	0.42200	0.01560	10	-28.8430	0.044900	0.993000	0.919000	0.10600	0.24800	115.4160	148299	4
A quiet ...	Christia...	0.45700	0.00836	7	-28.7330	0.039500	0.995000	0.955000	0.08570	0.28500	54.6210	122445	4
A tale t...	Luiza Sc...	0.38600	0.01280	1	-28.6600	0.039600	0.991000	0.931000	0.10900	0.18900	126.3740	130134	4
ATLiens	Outkast	0.91800	0.73400	11	-2.8320	0.269000	0.029600	0.000008	0.19100	0.60800	97.0440	230693	4
Adieux	Ludovico...	0.22500	0.00407	3	-37.1070	0.046400	0.987000	0.936000	0.11200	0.32000	73.5950	175493	4
Afterlig...	Arlo Thi...	0.38300	0.01430	1	-30.2790	0.044400	0.993000	0.914000	0.09750	0.18300	71.0290	200250	3
Ain't No...	Luke Com...	0.48700	0.65800	5	-9.9730	0.029600	0.008770	0.007770	0.10000	0.28200	142.1800	210950	4
Almenno ...	Rhian Ca...	0.43400	0.01200	2	-35.8260	0.032100	0.992000	0.896000	0.10600	0.29300	107.4370	167189	4
Almost A...	Florenti...	0.43000	0.06880	7	-27.3940	0.026800	0.990000	0.909000	0.10700	0.21400	83.5730	173133	4
Am I Oka...	Megan Mo...	0.59300	0.73400	9	-5.4160	0.051400	0.020000	0.000000	0.13800	0.51800	125.9370	235003	4
Ante Up ...	M.O.P.	0.69900	0.79300	1	-4.8560	0.268000	0.005140	0.000003	0.70000	0.92900	94.1380	248693	4
Arabesco	Lorenzo ...	0.36800	0.00394	3	-34.8920	0.050600	0.991000	0.903000	0.11400	0.13900	103.9680	155944	4
Arbor	Samuel K...	0.35200	0.00519	2	-34.3590	0.043600	0.995000	0.963000	0.09810	0.41400	89.4750	155500	4

...

Spotify API: 250 instances (n=250), 12 variables (p=12)

Data cloud

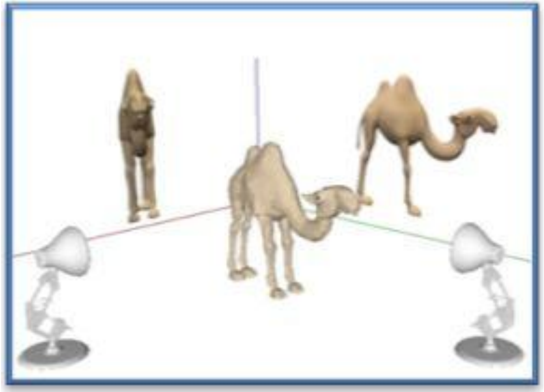
instances

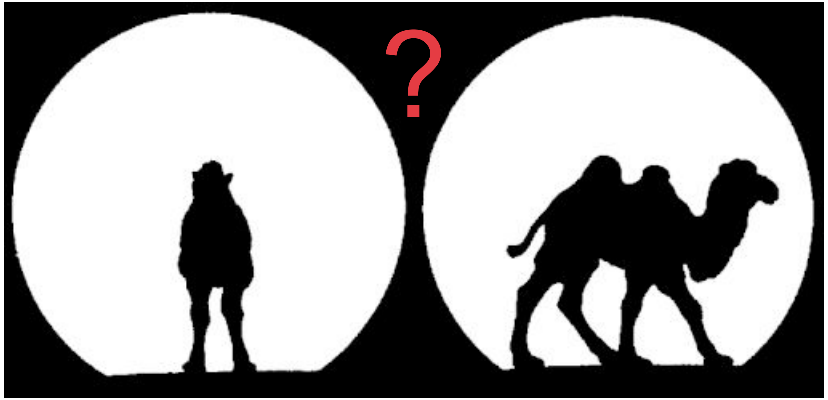
Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

Instance: $\boldsymbol{x}_i = (x_i^1, ..., x_i^p)$

Data cloud: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

<u>Objective:</u> provide the best **simplified** representation of the data





Example 3D→2D: which projection seems better?

Data cloud

instances

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

Instance: $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^p)$

Data cloud: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

<u>Objective:</u> provide the best **simplified** representation of the data

- Find a subspace $\boldsymbol{E_k}$ of $\mathbb{R}^p$ of dimension k
- Define k **new** variables linear combination of the p initial variables
- Lose as little information as possible

New variables: **principal components**

Data cloud

instances

Name	bitter	sweet	acid	salted	alcaline
St Yorre	3.4	3.1	2.9	6.4	4.8
Banoit	3.8	2.6	2.7	4.7	4.5
Vichy	2.9	2.9	2.1	6.0	5.0
Quézac	3.9	2.6	3.8	4.7	4.3

Instance: $\boldsymbol{x}_i = (x_i^1, \dots, x_i^p)$

Data cloud: $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$

<u>Objective:</u> provide the best **simplified** representation of the data

- Find a subspace $\boldsymbol{E_k}$ of $\mathbb{R}^\mathrm{p}$ of dimension k
- Define k new variables linear combination of the p initial variables
- Lose as little information as possible

New variables: **principal components**

Data cloud

"Lose as little information as possible"

- Find the best $\boldsymbol{E_k}$ that fits as well as possible to the cloud in instances
- The sum of of the squared **distances** of the instances to $\boldsymbol{E_k}$ must be minimized
- Subspace $\boldsymbol{E_k}$ where the projected cloud of instances has maximum **inertia**

Data cloud

"Lose as little information as possible"

- Find the best $\boldsymbol{E_k}$ that fits as well as possible to the cloud in instances
- The sum of of the squared **distances** of the instances to $\boldsymbol{E_k}$ must be minimized
- Subspace $\boldsymbol{E_k}$ where the projected cloud of instances has maximum **inertia**

Euclidean distance in a space of different units?

Inertia?

Center, scale, and standardize the data

Centering: subtracting the mean value of the **variable**

$$x_i^j \leftarrow x_i^j - \mu_j \text{ where } \mu_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Scaling: dividing by the standard deviation of the **variable**

$$x_i^j \leftarrow x_i^j / \sigma_j \text{ where } \sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^j - \mu_j)^2$$

Standardization: centering and scaling

$$x_i^j \leftarrow \frac{x_i^j - \mu^j}{\sigma^j}$$

Standardized data

We set:

$$\tilde{x}_i^j = \frac{x_i^j - \mu^j}{\sigma_j}$$

Standardized variable:

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^p \\ \vdots & \ddots & \vdots \\ \tilde{x}_n^1 & \dots & \tilde{x}_n^p \end{pmatrix}$$

Data

name	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	sign.
190306-1...	Hideyuki...	0.43100	0.02030	9	-27.7570	0.048400	0.993000	0.940000	0.09060	0.26600	175.4410	155533	3
1st of T...	Bone Thu...	0.72900	0.58100	5	-8.2350	0.180000	0.077700	0.000004	0.69600	0.50800	74.0380	314680	4
28	Zach Bry...	0.49200	0.51900	7	-6.8860	0.028800	0.227000	0.000085	0.06930	0.43500	80.8680	233333	3
93 ’Til ...	Souls Of...	0.59000	0.67200	1	-11.7920	0.412000	0.125000	0.000001	0.14700	0.68800	206.2470	286440	4
A Bar So...	Shabooze...	0.72200	0.70900	9	-4.9500	0.027300	0.063300	0.000000	0.08040	0.60400	81.0120	171292	4
A Cigare...	Gavin Ad...	0.61000	0.15000	7	-11.9460	0.028400	0.855000	0.000002	0.12800	0.21000	136.0890	179883	4
A New St...	Ferragno	0.35000	0.01500	8	-28.4700	0.050900	0.994000	0.958000	0.11500	0.29300	124.7840	155040	4
A Safe S...	Aramis M...	0.42200	0.01560	10	-28.8430	0.044900	0.993000	0.919000	0.10600	0.24800	115.4160	148299	4
A quiet ...	Christia...	0.45700	0.00836	7	-28.7330	0.039500	0.995000	0.955000	0.08570	0.28500	54.6210	122445	4
A tale t...	Luiza Sc...	0.38600	0.01280	1	-28.6600	0.039600	0.991000	0.931000	0.10900	0.18900	126.3740	130134	4
ATLiens	Outkast	0.91800	0.73400	11	-2.8320	0.269000	0.029600	0.000008	0.19100	0.60800	97.0440	230693	4
Adieux	Ludovico...	0.22500	0.00407	3	-37.1070	0.046400	0.987000	0.936000	0.11200	0.32000	73.5950	175493	4
Afterlig...	Arlo Thi...	0.38300	0.01430	1	-30.2790	0.044400	0.993000	0.914000	0.09750	0.18300	71.0290	200250	3
Ain’t No...	Luke Com...	0.48700	0.65800	5	-9.9730	0.029600	0.008770	0.007770	0.10000	0.28200	142.1800	210950	4
Almenno ...	Rhian Ca...	0.43400	0.01200	2	-35.8260	0.032100	0.992000	0.896000	0.10600	0.29300	107.4370	167189	4
Almost A...	Florenti...	0.43000	0.06880	7	-27.3940	0.026800	0.990000	0.909000	0.10700	0.21400	83.5730	173133	4
Am I Oka...	Megan Mo...	0.59300	0.73400	9	-5.4160	0.051400	0.020000	0.000000	0.13800	0.51800	125.9370	235003	4
Ante Up ...	M.O.P.	0.69900	0.79300	1	-4.8560	0.268000	0.005140	0.000003	0.70000	0.92900	94.1380	248693	4
Arabesco	Lorenzo ...	0.36800	0.00394	3	-34.8920	0.050600	0.991000	0.903000	0.11400	0.13900	103.9680	155944	4
Arbor	Samuel K...	0.35200	0.00519	2	-34.3590	0.043600	0.995000	0.963000	0.09810	0.41400	89.4750	155500	4

...

Standardized data

name	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	sign.
190306-1...	Hideyuki...	-0.8549	-1.1562	1.0508	-0.9291	-0.6585	1.1639	1.2381	-0.5729	-0.8216	2.2229	-0.8551	-2.4567
1st of T...	Bone Thu...	0.7435	0.4922	-0.0439	0.6791	0.4271	-0.9308	-0.8336	3.1925	0.1860	-1.0725	1.9728	0.2620
28	Zach Bry...	-0.5277	0.3099	0.5034	0.7903	-0.8202	-0.5891	-0.8334	-0.7053	-0.1178	-0.8505	0.5273	-2.4567
93 'Til ...	Souls Of...	-0.0020	0.7597	-1.1388	0.3861	2.3411	-0.8226	-0.8336	-0.2221	0.9356	3.2240	1.4710	0.2620
A Bar So...	Shabooze...	0.7060	0.8685	1.0508	0.9498	-0.8326	-0.9638	-0.8336	-0.6363	0.5858	-0.8458	-0.5750	0.2620
A Cigare...	Gavin Ad...	0.1052	-0.7748	0.5034	0.3734	-0.8235	0.8481	-0.8336	-0.3402	-1.0548	0.9440	-0.4224	0.2620
A New St...	Ferragno	-1.2893	-1.1717	0.7771	-0.9878	-0.6379	1.1662	1.2778	-0.4211	-0.7091	0.5766	-0.8638	0.2620
A Safe S...	Aramis M...	-0.9031	-1.1700	1.3246	-1.0185	-0.6874	1.1639	1.1918	-0.4771	-0.8965	0.2721	-0.9836	0.2620
A quiet ...	Christia...	-0.7154	-1.1913	0.5034	-1.0095	-0.7319	1.1685	1.2712	-0.6033	-0.7425	-1.7035	-1.4430	0.2620
A tale t...	Luiza Sc...	-1.0962	-1.1782	-1.1388	-1.0035	-0.7311	1.1593	1.2183	-0.4584	-1.1422	0.6283	-1.3064	0.2620
ATLiens	Outkast	1.7573	0.9420	1.5983	1.1243	1.1613	-1.0409	-0.8336	0.0515	0.6024	-0.3248	0.4804	0.2620
Adieux	Ludovico...	-1.9598	-1.2039	-0.5913	-1.6994	-0.6750	1.1502	1.2293	-0.4398	-0.5967	-1.0869	-0.5004	0.2620
Afterlig...	Arlo Thi...	-1.1123	-1.1738	-1.1388	-1.1368	-0.6915	1.1639	1.1808	-0.5299	-1.1672	-1.1703	-0.0605	-2.4567
Ain't No...	Luke Com...	-0.5545	0.7186	-0.0439	0.5359	-0.8136	-1.0886	-0.8165	-0.5144	-0.7550	1.1419	0.1296	0.2620
Almenno ...	Rhian Ca...	-0.8388	-1.1806	-0.8651	-1.5938	-0.7930	1.1616	1.1412	-0.4771	-0.7091	0.0128	-0.6479	0.2620
Almost A...	Florenti...	-0.8602	-1.0136	0.5034	-0.8992	-0.8367	1.1571	1.1698	-0.4709	-1.0381	-0.7626	-0.5423	0.2620
Am I Oka...	Megan Mo...	0.0140	0.9420	1.0508	0.9114	-0.6338	-1.0629	-0.8336	-0.2780	0.2277	0.6141	0.5570	0.2620
Ante Up ...	M.O.P.	0.5826	1.1155	-1.1388	0.9575	1.1531	-1.0969	-0.8336	3.2174	1.9391	-0.4193	0.8002	0.2620
Arabesco	Lorenzo ...	-1.1928	-1.2043	-0.5913	-1.5169	-0.6404	1.1593	1.1566	-0.4273	-1.3504	-0.0998	-0.8478	0.2620
Arbor	Samuel K...	-1.2786	-1.2006	-0.8651	-1.4730	-0.6981	1.1685	1.2888	-0.5262	-0.2053	-0.5708	-0.8556	0.2620

...

Standardized data: $\forall\ j\ ,\ \mu^j = 0$ and $\sigma^j = 1$

Standardized data

When should data be standardized ?

- **Essential** when variables are not expressed in the same unit.

- Generally recommended: gives **equal importance** to each variable.

- Huge influence on study results ⚠️ ⚠️

→ To do almost all the time…

Standardized data with weights

It can be useful to weight instances

Each instance i is associated with a weight w_i such as:

$$\forall i, w_i \geq 0 \text{ and } \sum_{i=1}^{n} w_i = 1$$

Without weighting:

$$w_i = 1/n$$

Inertia

Total inertia: **dispersion** of the data cloud around the barycenter

$$I = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 \text{ with } \boldsymbol{\mu} = (\mu^1, ..., \mu^p)$$

For standardized data,

$$\boldsymbol{\mu} = \vec{0} \text{ and } \|\boldsymbol{x}_i\|^2 = \sum_{j=1}^{p} (\boldsymbol{x}_i^j)^2 \text{ so } I = p$$

Weighted inertia:

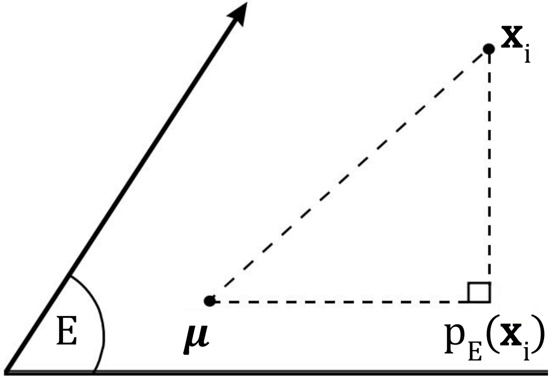
$$I = \sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 \text{ with } \boldsymbol{\mu} = (\mu^1, ..., \mu^p) \text{ and } \mu^j = \frac{1}{n}\sum_{i=1}^{n} w_i x_i^j$$

Equivalence distance/inertia

Inertia of the projection onto a subspace E where the data is projected

$$I_E = \frac{1}{n} \sum_{i=1}^{n} \|p_E(\boldsymbol{x}_i) - \boldsymbol{\mu}\|^2$$

where $p_E(\boldsymbol{x}_i)$ is the orthogonal projection of point $\boldsymbol{x}_i$ onto the subspace E



$$\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 = \|\boldsymbol{x}_i - p_E(\boldsymbol{x}_i)\|^2 + \|p_E(\boldsymbol{x}_i) - \boldsymbol{\mu}\|^2$$

total inertia

minimize distances

maximize inertia of the projection

Maximizing inertia

<u>Question:</u> How to determine the optimal subspace E that maximizes the inertia?

- Find new axes with maximum inertia
- Change of basis in $\mathbb{R}^{p}$ where the first axis has maximum possible

Let $\boldsymbol{V} \in \mathbb{R}^{pK}$ be the projection matrix with $\mathbf{v_k}$ being the k-th axis.

Let $\boldsymbol{XV} \in \mathbb{R}^{nK}$ be projection of the data $\boldsymbol{X}$ onto the subspace defined by $\boldsymbol{V}$

$$\text{Inertia}(\boldsymbol{XV}) = \frac{1}{n}\sum_{i=1}^{n} \|(\boldsymbol{XV})_i\|^2 = \frac{1}{n}\langle \boldsymbol{XV}, \boldsymbol{XV}\rangle = \text{trace}(\boldsymbol{V}^T \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V})$$

Covariance matrix

Maximizing inertia

Covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} \sigma_1^2 & \mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^2) & \dots & \mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^p) \\ \mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^2) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^p) & \dots & \dots & \sigma_p^2 \end{pmatrix}$$

where $\mathrm{Cov}(\boldsymbol{x}^j,\boldsymbol{x}^{j'}) = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \mu^j)(x_i^{j'} - \mu^{j'})$ and $\sigma_j^2 = \mathrm{Cov}(\boldsymbol{x}^j,\boldsymbol{x}^j)$

Correlation matrix:

$$\boldsymbol{C} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & \frac{\mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^2)}{\sigma_1\sigma_2} & \dots & \frac{\mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^p)}{\sigma_1\sigma_p} \\ \frac{\mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^2)}{\sigma_1\sigma_2} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\mathrm{Cov}(\boldsymbol{x}^1,\boldsymbol{x}^p)}{\sigma_1\sigma_p} & \dots & \dots & 1 \end{pmatrix}$$

Values in [-1;1]
→ interpretable

Maximizing inertia

Question: How to find $\boldsymbol{V}$?

Since correlation matrix $\boldsymbol{C}$ is symmetric: $\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$

- $\boldsymbol{Q} \in \mathbb{R}^{p \times p}$ orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{C}$
- $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times p}$ diagonal matrix with the corresponding eigenvalues $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p$ of $\boldsymbol{C}$

Let's write $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{U}$

Thus,

$$\mathrm{trace}(\boldsymbol{V}^T\boldsymbol{C}\boldsymbol{V}) = \mathrm{trace}((\boldsymbol{Q}\boldsymbol{U})^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{U})) = \mathrm{trace}(\boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U})$$

The trace is maximal when $\boldsymbol{U} = \boldsymbol{I}$ so $\boxed{\boldsymbol{V} = \boldsymbol{Q}}$

PCA

Points of view:

- **Geometric**: directions of maximum inertia
- **Statistical**: independent axes that best explain the variance of the data

How it works:

- Standardize the data
- Compute the covariance matrix
- Extract eigenvalues and eigenvectors
- Project the data onto the principal components

Eigenvalues, eigenvectors

Eigenvalues: $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p \times p}$ with $\lambda_1 \geq \cdots \geq \lambda_p$

Eigenvectors: $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) \in \mathbb{R}^{p \times p} \quad \rightarrow \quad \boldsymbol{V}_K = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) \in \mathbb{R}^{p \times K}$

Data projection: $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{V}_K \in \mathbb{R}^{n \times K}$

Principal components: $\boldsymbol{s}^k = (s_1^k, \ldots, s_n^k)^T$

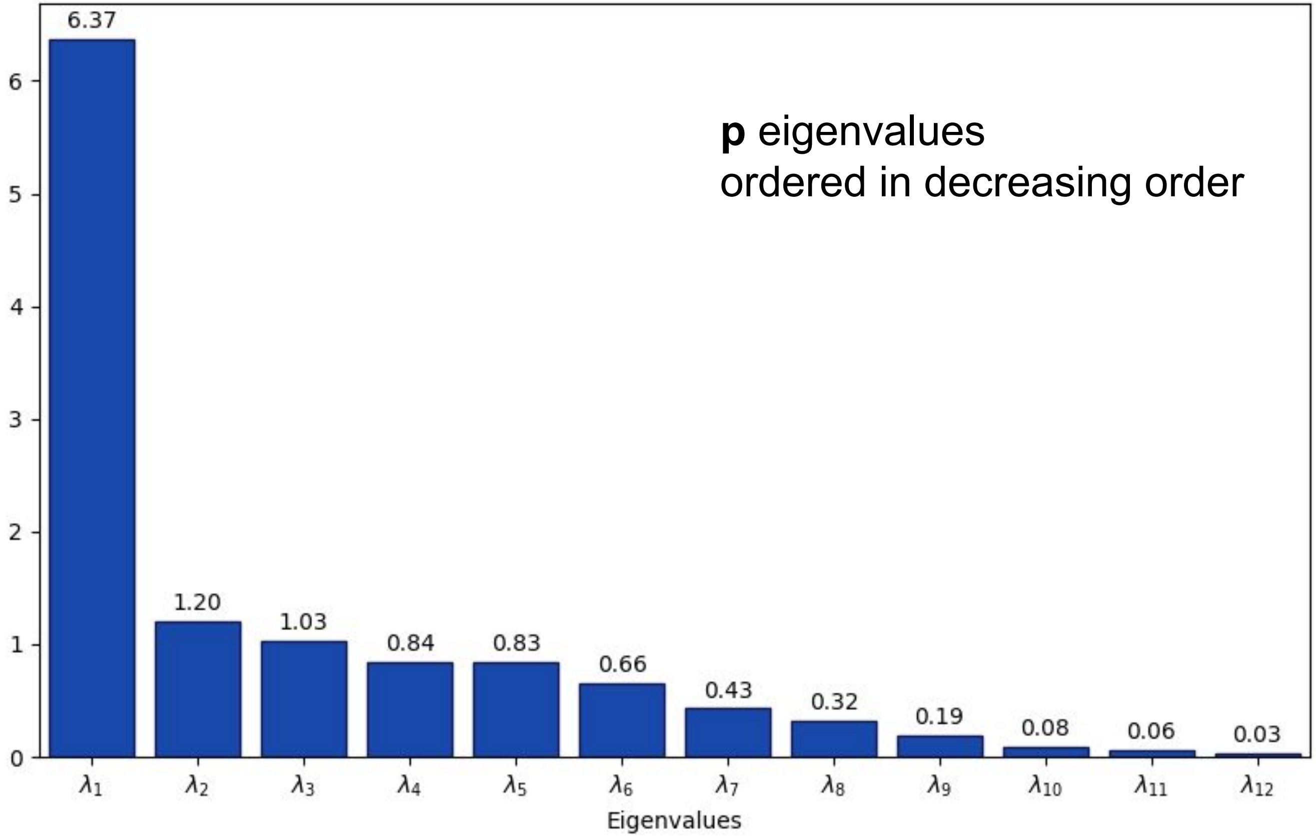
Standardized data

name	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration	sign.
190306-1...	Hideyuki...	-0.8549	-1.1562	1.0508	-0.9291	-0.6585	1.1639	1.2381	-0.5729	-0.8216	2.2229	-0.8551	-2.4567
1st of T...	Bone Thu...	0.7435	0.4922	-0.0439	0.6791	0.4271	-0.9308	-0.8336	3.1925	0.1860	-1.0725	1.9728	0.2620
28	Zach Bry...	-0.5277	0.3099	0.5034	0.7903	-0.8202	-0.5891	-0.8334	-0.7053	-0.1178	-0.8505	0.5273	-2.4567
93 'Til ...	Souls Of...	-0.0020	0.7597	-1.1388	0.3861	2.3411	-0.8226	-0.8336	-0.2221	0.9356	3.2240	1.4710	0.2620
A Bar So...	Shabooze...	0.7060	0.8685	1.0508	0.9498	-0.8326	-0.9638	-0.8336	-0.6363	0.5858	-0.8458	-0.5750	0.2620
A Cigare...	Gavin Ad...	0.1052	-0.7748	0.5034	0.3734	-0.8235	0.8481	-0.8336	-0.3402	-1.0548	0.9440	-0.4224	0.2620
A New St...	Ferragno	-1.2893	-1.1717	0.7771	-0.9878	-0.6379	1.1662	1.2778	-0.4211	-0.7091	0.5766	-0.8638	0.2620
A Safe S...	Aramis M...	-0.9031	-1.1700	1.3246	-1.0185	-0.6874	1.1639	1.1918	-0.4771	-0.8965	0.2721	-0.9836	0.2620
A quiet ...	Christia...	-0.7154	-1.1913	0.5034	-1.0095	-0.7319	1.1685	1.2712	-0.6033	-0.7425	-1.7035	-1.4430	0.2620
A tale t...	Luiza Sc...	-1.0962	-1.1782	-1.1388	-1.0035	-0.7311	1.1593	1.2183	-0.4584	-1.1422	0.6283	-1.3064	0.2620
ATLiens	Outkast	1.7573	0.9420	1.5983	1.1243	1.1613	-1.0409	-0.8336	0.0515	0.6024	-0.3248	0.4804	0.2620
Adieux	Ludovico...	-1.9598	-1.2039	-0.5913	-1.6994	-0.6750	1.1502	1.2293	-0.4398	-0.5967	-1.0869	-0.5004	0.2620
Afterlig...	Arlo Thi...	-1.1123	-1.1738	-1.1388	-1.1368	-0.6915	1.1639	1.1808	-0.5299	-1.1672	-1.1703	-0.0605	-2.4567
Ain't No...	Luke Com...	-0.5545	0.7186	-0.0439	0.5359	-0.8136	-1.0886	-0.8165	-0.5144	-0.7550	1.1419	0.1296	0.2620
Almenno ...	Rhian Ca...	-0.8388	-1.1806	-0.8651	-1.5938	-0.7930	1.1616	1.1412	-0.4771	-0.7091	0.0128	-0.6479	0.2620
Almost A...	Florenti...	-0.8602	-1.0136	0.5034	-0.8992	-0.8367	1.1571	1.1698	-0.4709	-1.0381	-0.7626	-0.5423	0.2620
Am I Oka...	Megan Mo...	0.0140	0.9420	1.0508	0.9114	-0.6338	-1.0629	-0.8336	-0.2780	0.2277	0.6141	0.5570	0.2620
Ante Up ...	M.O.P.	0.5826	1.1155	-1.1388	0.9575	1.1531	-1.0969	-0.8336	3.2174	1.9391	-0.4193	0.8002	0.2620
Arabesco	Lorenzo ...	-1.1928	-1.2043	-0.5913	-1.5169	-0.6404	1.1593	1.1566	-0.4273	-1.3504	-0.0998	-0.8478	0.2620
Arbor	Samuel K...	-1.2786	-1.2006	-0.8651	-1.4730	-0.6981	1.1685	1.2888	-0.5262	-0.2053	-0.5708	-0.8556	0.2620

...

$\tilde{X}$

Eigenvalues



p eigenvalues
ordered in decreasing order

Variance explained

λ_j : inertia of the data cloud projected onto the j-th axis
variance explained by the j-th axis

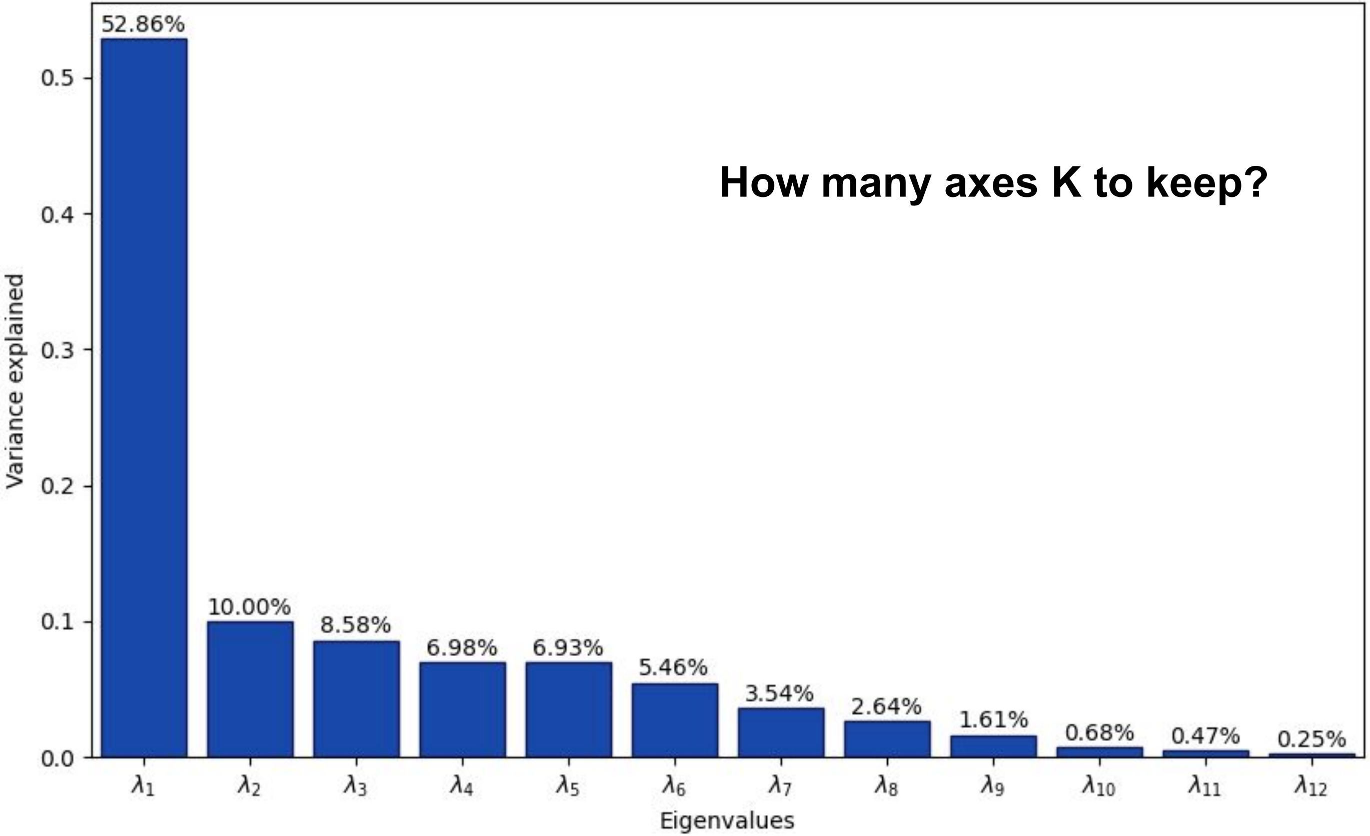
$I_{E_K} = \lambda_1 + ... + \lambda_K$: inertia of the data cloud projected onto the subspace $\boldsymbol{E_K}$
variance explained by the K first axes

$I = \lambda_1 + ... + \lambda_p$: total inertia

Proportion of variance explained by the first K axes:

$$\frac{I_{E_K}}{I} = \frac{\lambda_1 + ... + \lambda_K}{\lambda_1 + ... + \lambda_p}$$

Eigenvalues

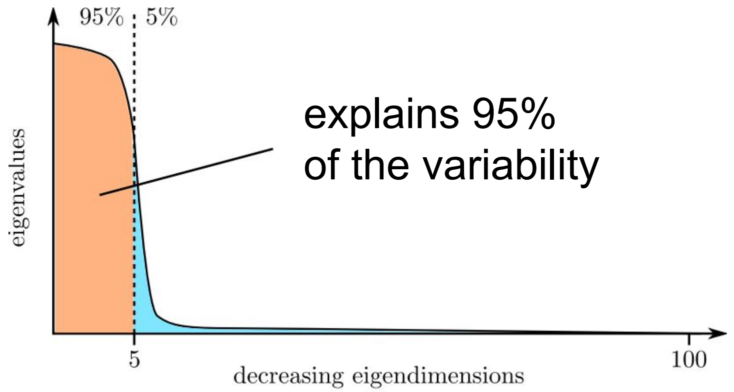


How many axes K to keep?

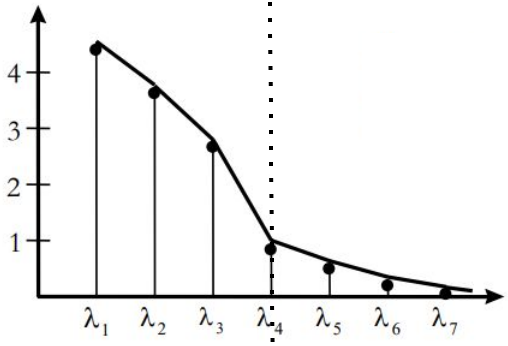
Number of axes to keep

Two different criteria:

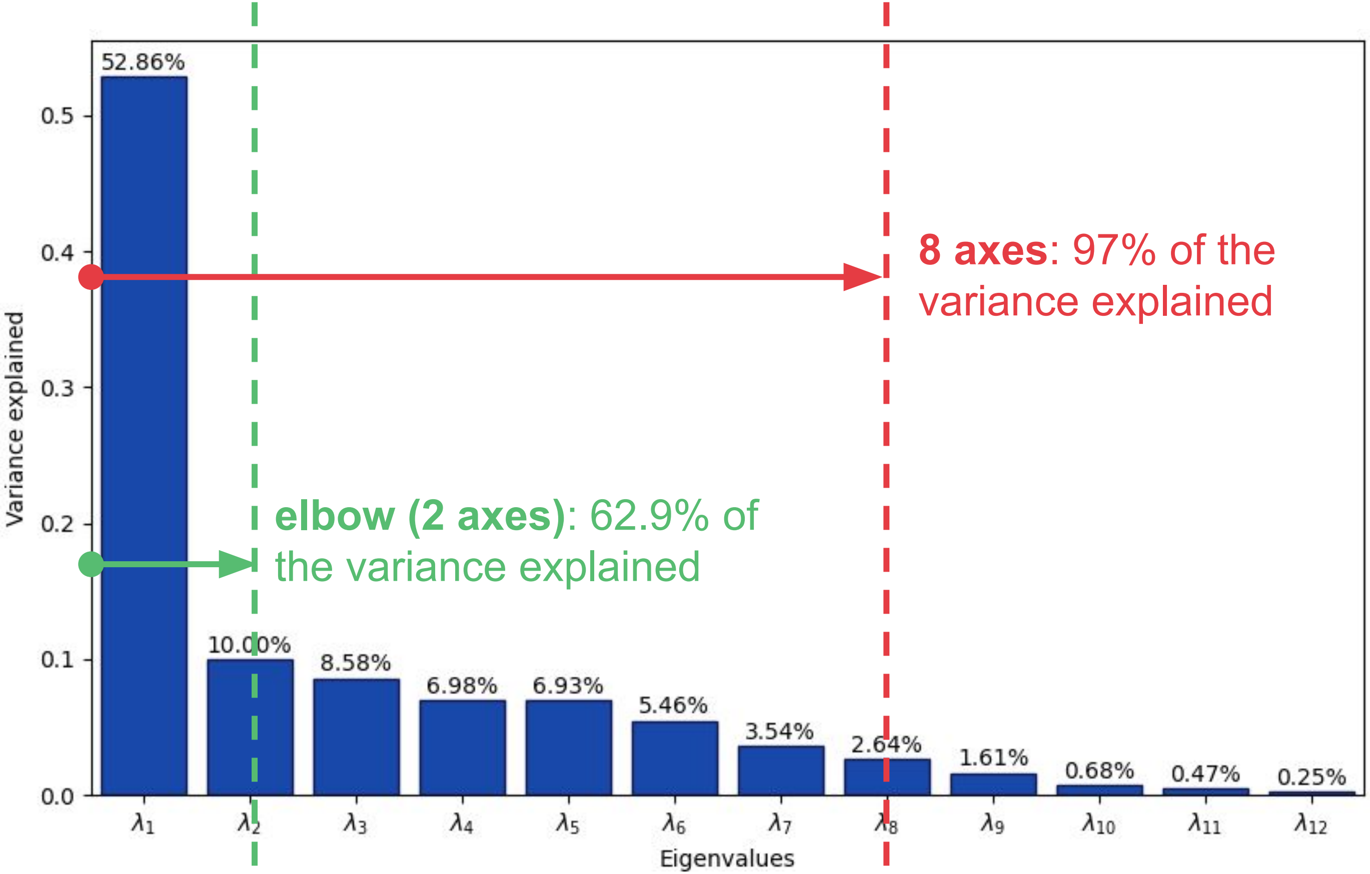
- Keep axes that explain 95% of the variance



- Heuristic: elbow criterion



Eigenvalues



Projected data

name	artist	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
190306-1...	Hideyuki...	2.99084	2.45148	-1.27358	-1.42516	0.05321	0.94915	0.17419	-0.08759	0.63915	0.06993	0.22102	0.23690
1st of T...	Bone Thu...	-2.75400	-0.95215	-0.70456	0.22152	2.55414	-0.99003	-1.17355	-0.61950	-0.01484	-0.01678	-0.18094	0.07622
28	Zach Bry...	-0.24675	0.37198	-2.11362	-1.32765	-1.26577	-1.21513	-0.44531	-0.13778	-0.71184	0.07937	-0.52654	-0.00507
93 'Til ...	Souls Of...	-2.44318	1.39613	1.89437	-2.00597	0.34621	2.54613	-0.19956	0.23083	-0.36362	-0.15316	-0.21022	0.03277
A Bar So...	Shabooze...	-1.30580	0.28978	-0.53976	1.05513	-1.51418	-1.27779	0.50472	0.04544	0.16673	-0.17879	0.00271	-0.06192
A Cigare...	Gavin Ad...	0.81814	1.08537	0.37003	0.32680	-0.46867	-0.18956	-0.71258	-0.03709	0.98116	1.12227	-0.55942	-0.26979
A New St...	Ferragno	2.83394	0.85882	0.13329	0.78288	0.25100	0.52095	0.05022	-0.09341	-0.19430	0.04524	-0.01947	0.24076
A Safe S...	Aramis M...	2.78061	0.84572	-0.31932	1.25253	0.06509	0.45903	-0.03546	0.02169	0.11883	0.04938	0.04377	0.11667
A quiet ...	Christia...	2.94637	-1.02783	-0.60152	1.42600	-0.21980	-0.40071	0.46252	0.33304	-0.11616	0.06315	-0.00810	0.07368
A tale t...	Luiza Sc...	3.14906	-0.04518	1.28803	-0.35910	0.24030	0.04714	0.07546	0.25214	0.25231	0.10040	0.03635	0.17753
ATLiens	Outkast	-2.84464	0.21927	-1.07406	1.05686	-0.59230	0.44997	0.02659	0.50706	0.61716	0.00755	0.26676	0.13645
Adieux	Ludovico...	3.26260	-1.02613	0.27551	0.38363	0.41953	-0.02722	0.02020	-0.22428	-1.19198	-0.31802	-0.37805	-0.23582
Afterlig...	Arlo Thi...	3.15221	-1.24534	-1.44618	-1.85160	0.04577	-0.33057	-0.38166	-0.11399	-0.44819	0.04750	-0.06218	-0.00986
Ain't No...	Luke Com...	-0.50113	1.36321	0.94152	-0.38713	-0.88727	-0.76744	-1.01276	0.14466	-0.13018	-0.46517	-0.21639	-0.04958
Almenno ...	Rhian Ca...	2.92487	-0.57001	0.75924	-0.04927	0.26915	0.28154	0.06183	-0.38013	0.06884	-0.21439	-0.03002	-0.27578
Almost A...	Florenti...	2.65495	-0.35204	-0.30018	0.99831	-0.02450	-0.05561	-0.35416	-0.14030	-0.19286	0.11910	0.11249	0.09410
Am I Oka...	Megan Mo...	-1.49962	1.37264	-0.03664	0.40976	-0.87311	-0.57035	-0.51554	-0.36214	-0.25574	-0.27599	-0.03859	0.06816
Ante Up ...	M.O.P.	-3.46229	-0.82485	0.32312	-0.61151	2.65642	-0.86265	0.93254	-0.15906	-0.39451	-0.03599	-0.01444	0.15477
Arabesco	Lorenzo ...	3.23485	-0.41720	0.62879	0.15431	0.33503	0.16268	-0.35336	0.16756	-0.05915	-0.15651	-0.06964	-0.19476
Arbor	Samuel K...	2.98830	-0.88722	0.59407	0.13281	0.24740	0.10261	0.58006	-0.32232	-0.52839	-0.19661	-0.16897	-0.09219

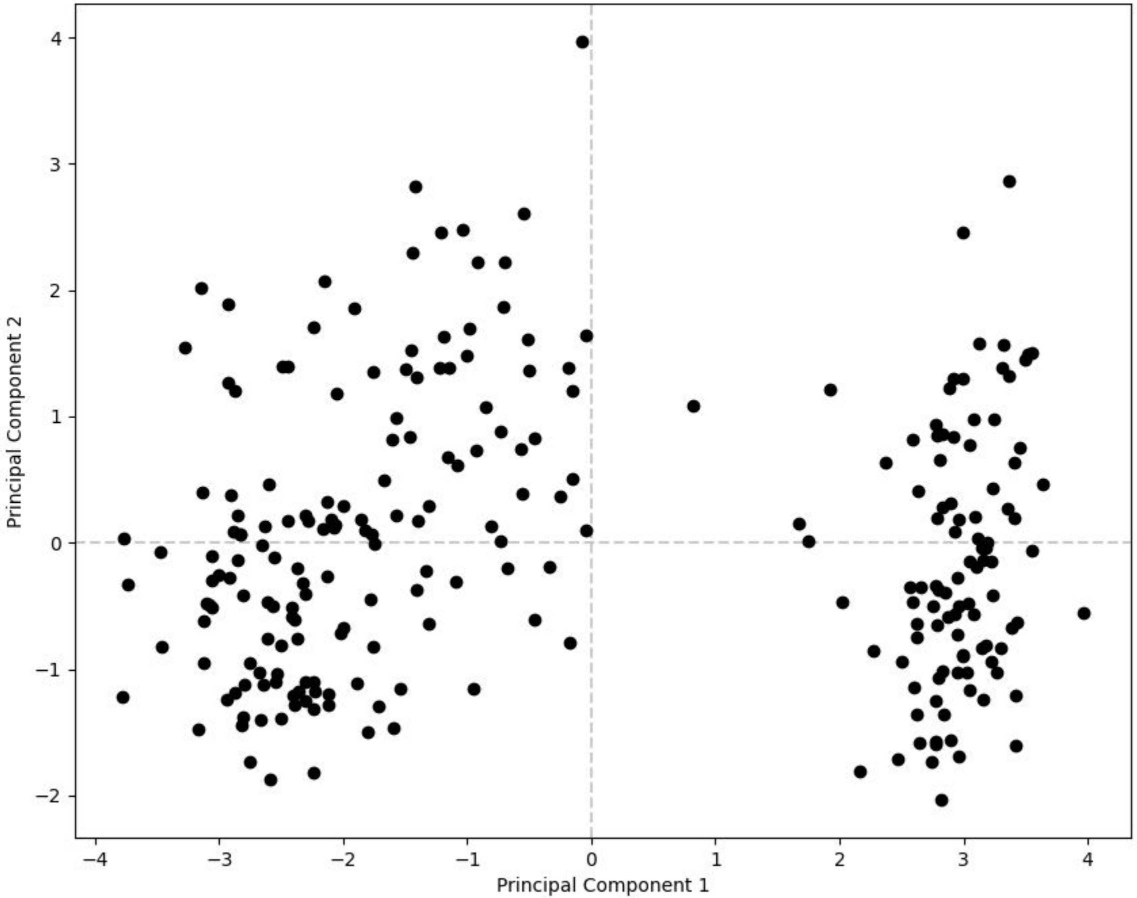
...

Projected data

name	artist	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
190306-1...	Hideyuki...	2.99084	2.45148	-1.27358	-1.42516	0.05321	0.94915	0.17419	-0.08759	0.63915	0.06993	0.22102	0.23690
1st of T...	Bone Thu...	-2.75400	-0.95215	-0.70456	0.22152	2.55414	-0.99003	-1.17355	-0.61950	-0.01484	-0.01678	-0.18094	0.07622
28	Zach Bry...	-0.24675	0.37198	-2.11362	-1.32765	-1.26577	-1.21513	-0.44531	-0.13778	-0.71184	0.07937	-0.52654	-0.00507
93 'Til ...	Souls Of...	-2.44318	1.39613	1.89437	-2.00597	0.34621	2.54613	-0.19956	0.23083	-0.36362	-0.15316	-0.21022	0.03277
A Bar So...	Shabooze...	-1.30580	0.28978	-0.53976	1.05513	-1.51418	-1.27779	0.50472	0.04544	0.16673	-0.17879	0.00271	-0.06192
A Cigare...	Gavin Ad...	0.81814	1.08537	0.37003	0.32680	-0.46867	-0.18956	-0.71258	-0.03709	0.98116	1.12227	-0.55942	-0.26979
A New St...	Ferragno	2.83394	0.85882	0.13329	0.78288	0.25100	0.52095	0.05022	-0.09341	-0.19430	0.04524	-0.01947	0.24076
A Safe S...	Aramis M...	2.78061	0.84572	-0.31932	1.25253	0.06509	0.45903	-0.03546	0.02169	0.11883	0.04938	0.04377	0.11667
A quiet ...	Christia...	2.94637	-1.02783	-0.60152	1.42600	-0.21980	-0.40071	0.46252	0.33304	-0.11616	0.06315	-0.00810	0.07368
A tale t...	Luiza Sc...	3.14906	-0.04518	1.28803	-0.35910	0.24030	0.04714	0.07546	0.25214	0.25231	0.10040	0.03635	0.17753
ATLiens	Outkast	-2.84464	0.21927	-1.07406	1.05686	-0.59230	0.44997	0.02659	0.50706	0.61716	0.00755	0.26676	0.13645
Adieux	Ludovico...	3.26260	-1.02613	0.27551	0.38363	0.41953	-0.02722	0.02020	-0.22428	-1.19198	-0.31802	-0.37805	-0.23582
Afterlig...	Arlo Thi...	3.15221	-1.24534	-1.44618	-1.85160	0.04577	-0.33057	-0.38166	-0.11399	-0.44819	0.04750	-0.06218	-0.00986
Ain't No...	Luke Com...	-0.50113	1.36321	0.94152	-0.38713	-0.88727	-0.76744	-1.01276	0.14466	-0.13018	-0.46517	-0.21639	-0.04958
Almenno ...	Rhian Ca...	2.92487	-0.57001	0.75924	-0.04927	0.26915	0.28154	0.06183	-0.38013	0.06884	-0.21439	-0.03002	-0.27578
Almost A...	Florenti...	2.65495	-0.35204	-0.30018	0.99831	-0.02450	-0.05561	-0.35416	-0.14030	-0.19286	0.11910	0.11249	0.09410
Am I Oka...	Megan Mo...	-1.49962	1.37264	-0.03664	0.40976	-0.87311	-0.57035	-0.51554	-0.36214	-0.25574	-0.27599	-0.03859	0.06816
Ante Up ...	M.O.P.	-3.46229	-0.82485	0.32312	-0.61151	2.65642	-0.86265	0.93254	-0.15906	-0.39451	-0.03599	-0.01444	0.15477
Arabesco	Lorenzo ...	3.23485	-0.41720	0.62879	0.15431	0.33503	0.16268	-0.35336	0.16756	-0.05915	-0.15651	-0.06964	-0.19476
Arbor	Samuel K...	2.98830	-0.88722	0.59407	0.13281	0.24740	0.10261	0.58006	-0.32232	-0.52839	-0.19661	-0.16897	-0.09219

...

Projected data



PCA summary

Key messages:

- PCA is a tool for visualizing multidimensional data (applicable in high-dimensional spaces) and for reducing dimensionality

- The low-dimensional space that best represents the data is determined by the eigenvectors of the correlation matrix (standardized PCA) or the variance-covariance matrix (non-standardized PCA)

- Eigenvalues represent the amount of information (variance) explained by each axis

⚠️ ⚠️ **Axes are not actual variables but linear combinations!** ⚠️ ⚠️

→ **Difficult to interpret!**

Correlation between components

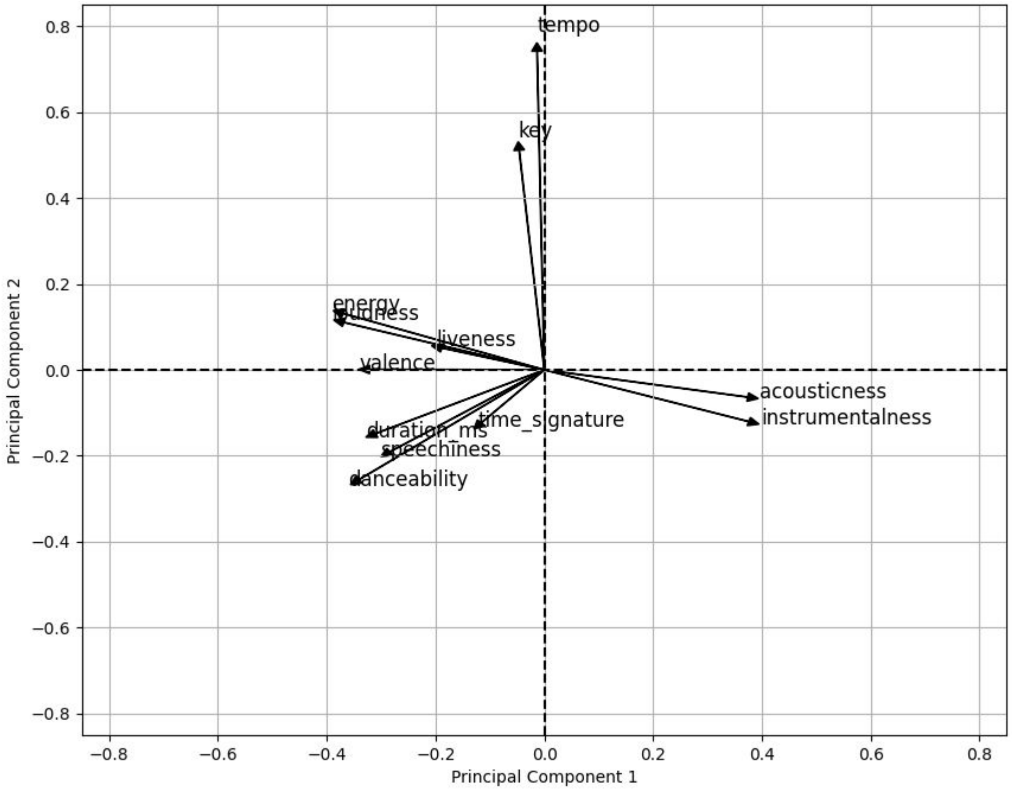
Correlation circle

Plot of each variable projected by $\boldsymbol{V}$: correlation circle

Variables with vectors close to the unit circle are well represented in the PC space.

Variables with collinear vectors are highly correlated.

Variables with vectors close to PC axes are correlated with new components.



Contributions

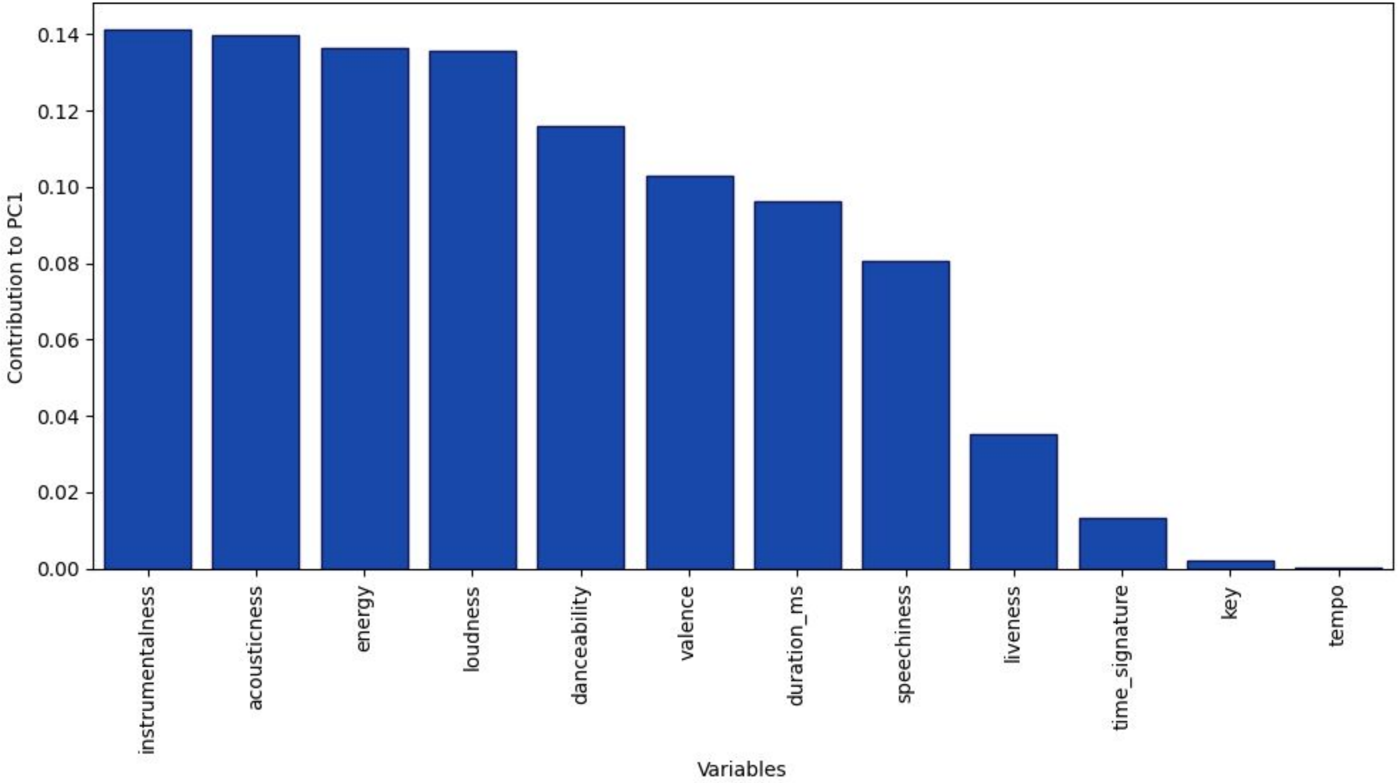
Contribution of instance i to the total inertia along the k-th component:

$$\mathrm{ctr}(i,k) = \frac{(s_i^k)^2}{\sum_{i'=1}^{n} (s_{i'}^k)^2}$$

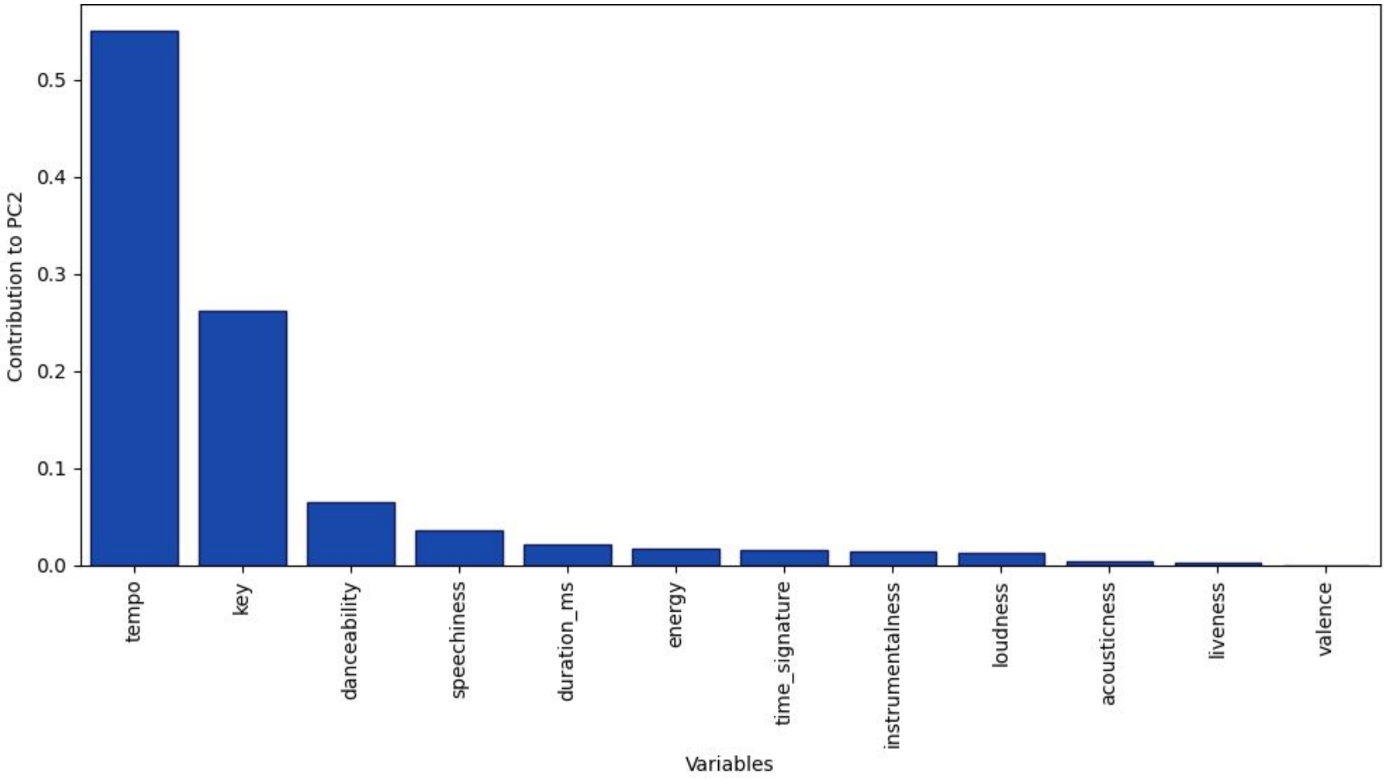
Quality of representation of instance i on axis k:

$$\mathrm{Q}(i,k) = \frac{(s_i^k)^2}{\sum_{j=1}^{p} (s_i^j)^2}$$

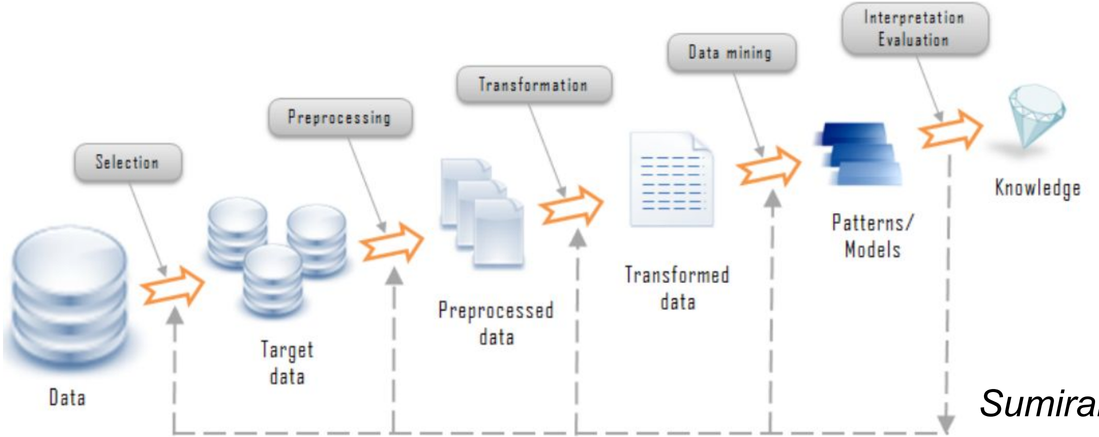
Contributions to PC1



Contributions to PC2



The data analysis process



Sumiran, 2018

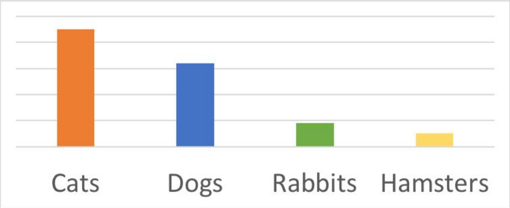
Data: songs and audio descriptors from SpotifyAPI
Selection: data from 250 songs
Preprocessing: Standardize data
Transformation: Compute the correlation matrix
Model: Eigenvectors, eigenvalues, projection
Knowledge: Correlation between variables (acousticness/instrumentalness, tempo/key)

Next course

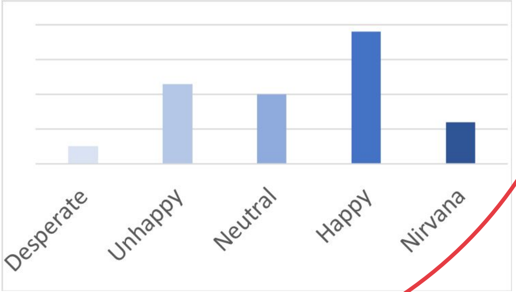
Types of data

Qualitative (categories)

Nominal:

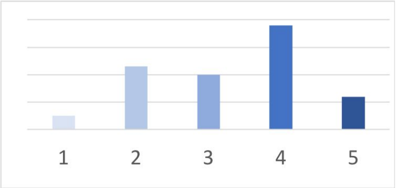


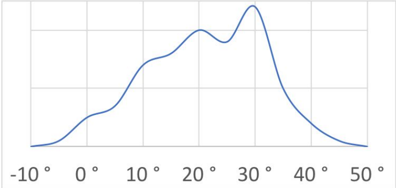
Ordered:



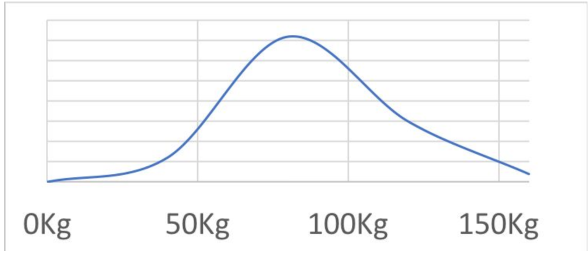
Quantitative (numerical values)

Interval (discrete or continuous):





Ratio:



Questions?

Sources, images courtesy and acknowledgment:
N. Papadakis, P-L Gonzales

Charles Brazier
charles.brazier@u-bordeaux.fr