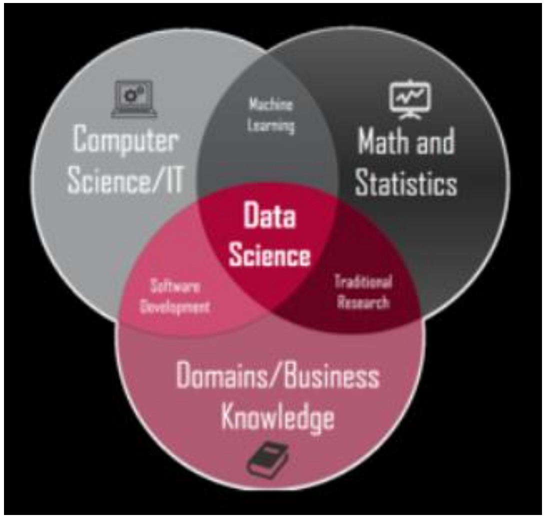
DAT200 - Applied machine learning

Subspace analysis

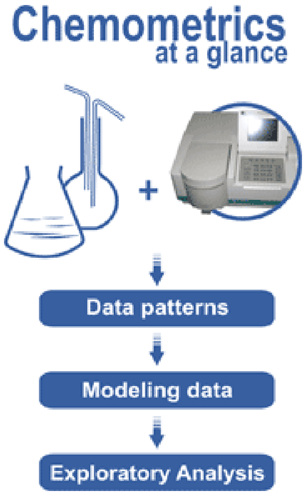
PCA, PCR, PLSR

Why Chemometrics in addition to ML?

- Wikipedia on **Data Science**
  - Data-driven science
  - Techniques and theories from mathematics, statistics, information science, and computer science



- Wikipedia on **Chemometrics**
  - Data-driven science
  - Methods from applied mathematics, multivariate statistics, and computer science



Why Chemometrics in addition to ML?

Chemometrics

- Languages
  - MATLAB, R
- Repositories
  - MATLAB file exchange, models.life.ku.dk, CRAN, …

Data Science

- Languages
  - Python, R, Java, Perl, C/C++
- Repositories
  - GitHub, CRAN, …

Why Chemometrics in addition to ML?

- Campus Ås has long and strong tradition in chemometrics
- Chemometrics has many similarities with ML
- Important **additional value** of using chemometrics
  - ML is mainly concerned with building good models for **prediction** of outcomes
  - ML is often considered as a **black box** (usually little interest in interpretation of variation in data)
  - Chemometrics is concerned with building good models that can be used for **interpretation** of the data and **prediction** of outcomes
  - Chemometrics handles well situations where data has many more (often highly correlated) variables than observations ($n << k$)

Why Chemometrics in addition to ML?

- Methods used in both chemometrics and ML
  - Principal component analysis (PCA)
  - Partial least squares regression (PLSR)
- Chemometrics methods we will discuss here
  - PCA, PLSR
  - Principal component regression (PCR)

Summary

- Subspace methods
  - Principal directions in data
  - Orthogonal components
  - Visualisation
  - Compression

- Predictor driven:
  - Principal component analysis (PCA)
  - Principal component regression (PCR)

- Response driven:
  - Partial least squares regression (PLSR)

Multivariate data – how to extract information?

OECD data: most frequent types (%) of cancer found in men from participating countries (Organisation for Economic Cooperation and Development)

	A	B	C	D	E	F	G	H	I	J	K
1	**MEN**	**Trachea-bronchus-lung**	**Colon, rectum and anus**	**Stomach**	**Pancreas**	**Prostate**	**Liver**	**Hodgkins disease**	**Leukemia**	**Bladder**	**Skin**
2	**Australia**	20.49342921	9.290023969	2.954789652	5.033473841	13.61269526	4.049921481	0.157037772	3.855690553	3.107694851	4.425985619
3	**Austria**	22.28692919	10.94472176	4.584951008	6.627842485	10.57496765	5.66648179	0.147901645	3.780735811	3.031983731	1.941209096
4	**Belgium**	30.09070593	10.45090049	3.253582227	5.179439989	9.162613382	3.450769029	0.144603655	3.56908111	4.719337452	1.018798475
5	**Canada**	27.73162434	11.44341588	3.011841654	5.419732574	9.739694596	3.863702297	0.195163119	3.789856792	3.534034866	1.5375689
6	**Chile**	13.304823	8.022491486	17.27251129	4.165676724	16.19545419	5.044745387	0.22174705	3.00942425	2.162033737	0.554367625
7	**Czech Rep.**	24.87532416	13.97034377	4.45508345	6.742469579	9.455415919	3.464326086	0.279273888	3.37123479	3.557417381	1.522707627
8	**Denmark**	24.31158521	12.16819648	3.212602332	6.375589184	14.30166212	3.162986852	0.21086579	3.100967502	4.366162243	2.245100471
9	**Estonia**	26.37195122	10.26422764	8.333333333	5.43699187	13.00813008	2.388211382	0.101626016	3.861788618	3.506097561	1.016260163
10	**Finland**	23.9178867	9.818586887	4.169318905	7.940802037	13.57415659	4.232972629	0.159134309	2.737110121	2.928071292	2.418841502
11	**France**	25.12575792	10.16168875	3.360655738	5.335728722	9.885470469	6.466427128	0.143723333	3.535818549	4.350999326	1.112732989
12	**Germany**	24.4012222	11.17718566	4.592272563	6.795183494	11.01291192	4.106843644	0.151953215	3.609915563	3.19840983	1.467785918
13	**Greece**	31.66724517	8.50399167	4.830498669	5.229665625	9.013074164	1.237996066	0.792548883	3.823903737	5.582552355	0.76940877
14	**Hungary**	30.41257367	16.08195341	5.287678922	5.29329217	6.797642436	3.104125737	0.117878193	2.559640752	3.575638507	1.044063991
15	**Iceland**	20.32258065	13.22580645	4.193548387	3.870967742	17.09677419	0.967741935	0.64516129	2.903225806	5.806451613	0.967741935
16	**Ireland**	22.95747731	12.32680363	4.395604396	5.805064501	12.4462494	3.511705686	0.238891543	3.057811753	2.866698519	1.887243192
17	**Israel**	21.82289737	12.67135976	5.446461652	8.725453872	7.725083364	3.501296777	0.351982216	5.094479437	4.890700259	2.278621712
18	**Italy**	26.06741808	10.90044415	6.087111372	5.398893824	7.628006369	6.95340652	0.250356155	3.660018436	4.687630939	1.131316517
19	**Japan**	23.99056121	11.97177581	14.73796762	7.315521922	5.327753633	9.132765224	0.051157496	2.214981311	2.426985349	0.151168096
20	**Korea**	26.20420989	10.09857518	13.10748569	5.630407645	3.142352891	18.28160647	0.094701046	1.977960484	1.975808187	0.271189359
21	**Luxembourg**	25.18382353	11.94852941	4.044117647	6.066176471	7.720588235	6.066176471	0.183823529	4.779411765	4.044117647	1.838235294
22	**Mexico**	11.49597401	6.91058059	8.227150728	5.088289306	16.32151434	7.591467721	0.740217545	6.260771295	1.873145925	0.802373217
23	**Netherlands**	27.15515358	11.46939311	3.543496308	5.334906279	11.07615677	2.127845502	0.157294534	3.062874121	3.700790842	2.154061257
24	**New Zealand**	19.7737655	13.09549706	4.198390255	4.763976506	12.72569067	3.56754405	0.195779856	4.15488362	2.74091799	5.286056124
25	**Norway**	21.39823009	13.45132743	3.203539823	6.194690265	17.48672566	2.566371681	0.17699115	3.221238938	4.247787611	3.309734513
26	**Poland**	30.65693431	11.9202253	6.439067379	4.550069927	8.201620783	2.0863268	0.19541353	2.927371305	5.145889611	1.41579018
27	**Portugal**	20.39060984	14.4667917	8.730518011	4.733880877	11.1039255	4.695078575	0.155209209	3.000711376	4.416995408	0.821315398
28	**Slovak Rep.**	22.66305123	15.09918653	6.108177537	5.651491366	7.606679035	3.11117454	0.285428857	2.554588269	3.125445983	1.541315827
29	**Slovenia**	24.94577007	13.72792067	7.468236752	5.11310815	11.09389526	3.997520917	0.154942671	2.479082739	4.09048652	2.169197397
30	**Spain**	26.77706347	14.07025989	5.245117455	4.827701776	8.81600195	5.138478413	0.19652052	2.953901466	6.384631791	0.885104049
31	**Sweden**	16.05304484	12.34514046	3.350200663	6.962135753	20.44145873	3.577037166	0.157040656	3.568312685	4.240097714	2.748211481
32	**Switzerland**	21.6075388	10.18847007	3.680709534	6.241685144	14.16851441	5.365853659	0	3.359201774	4.257206208	2.372505543
33	**Turkey**	38.97165809	7.623419473	8.846855339	5.211385946	7.225854048	3.723589565	0.219275775	3.612927024	3.397749862	0.57175646
34	**United Kingdom**	22.80308077	10.26261351	3.447751949	4.946063135	12.70759557	3.427883548	0.191671634	3.22335589	4.037960333	1.521685775
35	**United States**	29.1412459	9.06313206	2.226842334	6.211578252	9.487245059	4.537710188	0.237131309	4.268031445	3.463644848	1.993364309
36	**OECD**	25.99183401	10.69682199	6.27003433	5.933305685	9.235450758	5.686093446	0.194011091	3.432922396	3.617678042	1.304311271

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA)

What is it and for which situations can I use it for?

- Analysis of **one** data table $\boldsymbol{X}$
- Versatile method for almost all types of data to obtain an **overview** (for example in an early phase of investigation)
- **Explorative multivariate** statistical method
- Particularly suitable in situations …
  - With **lots** of data
  - Where **little** prior information is available
- Other names for PCA
  - Singular value decomposition (SVD)
  - Eigenvector decomposition

Principal Component Analysis (PCA)

How does it work and what kind of information do I get?

- Idea behind PCA – find the **most interesting** dimensions or directions of variability, so-called principal components
- Extracts **main information** (**systematic variation**) in the data
- Visualisation: present results graphically for interpretation
    - Information on **objects**
    - Information on **variables**
    - Other results

Principal Component Analysis (PCA)

What can I do with it / use it for?

- **Interpretation** of the **variance** in the data
  - Gain knowledge on how **objects** are distributed (patterns using background information)
  - Gain knowledge on how **variables** contribute to variance in the data
  - Generate hypotheses and ideas for further experimentation
- **Data pre-processing** and **data compression**
  - Use PCA as **filter** to get rid of **noise**
  - Use components **instead of** original data in subsequent analysis
  - **Dimensionality reduction** (as often used in ML) – use components as input in *classifiers*, *regression*, *clustering*, not original data
- **Classification** (not part of syllabus)
  - **SIMCA** method where PCA is applied for computations

PCA – data structure

1 K

1

	A	B	C	D	E	F	G	H	I	J	K
1	MEN	Trachea-bronchus-lung	Colon, rectum and anus	Stomach	Pancreas	Prostate	Liver	Hodgkins disease	Leukemia	Bladder	Skin
2	Australia	20.49342921	9.290023969	2.954789652	5.033473841	13.61269526	4.049921481	0.157037772	3.855690553	3.107694851	4.425985619
3	Austria	22.28692919	10.94472176	4.584951008	6.627842485	10.57496765	5.66648179	0.147901645	3.780735811	3.031983731	1.941209096
4	Belgium	30.09070593	10.45090049	3.253582227	5.179439989	9.162613382	3.450769029	0.144603655	3.56908111	4.719337452	1.018798475
5	Canada	27.73162434	11.44341588	3.011841654	5.419732574	9.739694596	3.863702297	0.195163119	3.789856792	3.534034866	1.5375689
6	Chile	13.304823	8.022491486	17.27251129	4.165676724	16.19545419	5.044745387	0.22174705	3.00942425	2.162033737	0.554367625
7	Czech Rep.	24.87532416	13.97034377	4.45508345	6.742469579	9.455415919	3.464326086	0.279273888	3.37123479	3.557417381	1.522707627
8	Denmark	24.31158521	12.16819648	3.212602332	6.375589184	14.30166212	3.162986852	0.21086579	3.100967502	4.366162243	2.245100471
9	Estonia	26.37195122	10.26422764	8.333333333	5.43699187	13.00813008	2.388211382	0.101626016	3.861788618	3.506097561	1.016260163
10	Finland	23.9178867	9.818586887	4.169318905	7.940802037	13.57415659	4.232972629	0.159134309	2.737110121	2.928071292	2.418841502
11	France	25.12575792	10.16168875	3.360655738	5.335728722	9.885470469	6.466427128	0.143723333	3.535818549	4.350999326	1.112732989
12	Germany	24.4012222	11.17718566	4.592272563	6.795183494	11.01291192	4.106843644	0.151953215	3.609915563	3.19840983	1.467785918
13	Greece	31.66724517	8.50399167	4.830498669	5.229665625	9.013074164	1.237996066	0.792548883	3.823903737	5.582552355	0.76940877
14	Hungary	30.41257367	16.08195341	5.287678922	5.29329217	6.797642436	3.104125737	0.117878193	2.559640752	3.575638507	1.044063991
15	Iceland	20.32258065	13.22580645	4.193548387	3.870967742	17.09677419	0.967741935	0.64516129	2.903225806	5.806451613	0.967741935
16	Ireland	22.95747731	12.32680363	4.395604396	5.805064501	12.4462494	3.511705686	0.238891543	3.057811753	2.866698519	1.887243192
17	Israel	21.82289737	12.67135976	5.446461652	8.725453872	7.725083364	3.501296777	0.351982216	5.094479437	4.890700259	2.278621712
18	Italy	26.06741808	10.90044415	6.087111372	5.398893824	7.628006369	6.95340652	0.250356155	3.660018436	4.687630939	1.131316517
19	Japan	23.99056121	11.97177581	14.73796762	7.315521922	5.327753633	9.132765224	0.051157496	2.214981311	2.426985349	0.151168096
20	Korea	26.20420989	10.09857518	13.10748569	5.630407645	3.142352891	18.28160647	0.094701046	1.977960484	1.975808187	0.271189359
21	Luxembourg	25.18382353	11.94852941	4.044117647	6.066176471	7.720588235	6.066176471	0.183823529	4.779411765	4.044117647	1.838235294
22	Mexico	11.49597401	6.91058059	8.227150728	5.088289306	16.32151434	7.591467721	0.740217545	6.260771295	1.873145925	0.802373217
23	Netherlands	27.15515358	11.46939311	3.543496308	5.334906279	11.07615677	2.127845502	0.157294534	3.062874121	3.700790842	2.154061257
24	New Zealand	19.7737655	13.09549706	4.198390255	4.763976506	12.72569067	3.56754405	0.195779856	4.15488362	2.74091799	5.286056124
25	Norway	21.39823009	13.45132743	3.203539823	6.194690265	17.48672566	2.566371681	0.17699115	3.221238938	4.247787611	3.309734513
26	Poland	30.65693431	11.9202253	6.439067379	4.550069927	8.201620783	2.0863268	0.19541353	2.927371305	5.145889611	1.41579018
27	Portugal	20.39060984	14.4667917	8.730518011	4.733880877	11.1039255	4.695078575	0.155209209	3.000711376	4.416995408	0.821315398
28	Slovak Rep.	22.66305123	15.09918653	6.108177537	5.651491366	7.606679035	3.11117454	0.285428857	2.554588269	3.125445983	1.541315827
29	Slovenia	24.94577007	13.72792067	7.468236752	5.11310815	11.09389526	3.997520917	0.154942671	2.479082739	4.09048652	2.169197397
30	Spain	26.77706347	14.07025989	5.245117455	4.827701776	8.81600195	5.138478413	0.19652052	2.953901466	6.384631791	0.885104049
31	Sweden	16.05304484	12.34514046	3.350200663	6.962135753	20.44145873	3.577037166	0.157040656	3.568312685	4.240097714	2.748211481
32	Switzerland	21.6075388	10.18847007	3.680709534	6.241685144	14.16851441	5.365853659	0	3.359201774	4.257206208	2.372505543
33	Turkey	38.97165809	7.623419473	8.846855339	5.211385946	7.225854048	3.723589565	0.219275775	3.612927024	3.397749862	0.57175646
34	United Kingdom	22.80308077	10.26261351	3.447751949	4.946063135	12.70759557	3.427883548	0.191671634	3.22335589	4.037960333	1.521685775
35	United States	29.1412459	9.06313206	2.226842334	6.211578252	9.487245059	4.537710188	0.237131309	4.268031445	3.463644848	1.993364309
36	OECD	25.99183401	10.69682199	6.27003433	5.933305685	9.235450758	5.686093446	0.194011091	3.432922396	3.617678042	1.304311271

N

$\boldsymbol{X}$

- Number of **objects (rows)**:
  - $n = 1 \dots N$
- Number of **variables (columns)**:
  - $k = 1 \dots K$
- Observed value x_{nk} for
  - n'th object
  - k'th variable

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NK} \end{pmatrix}$$

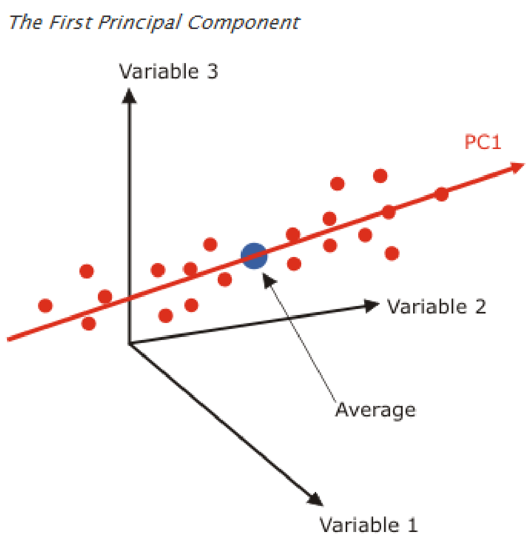
DEMO: OECD data – cancer in men

Concept behind PCA

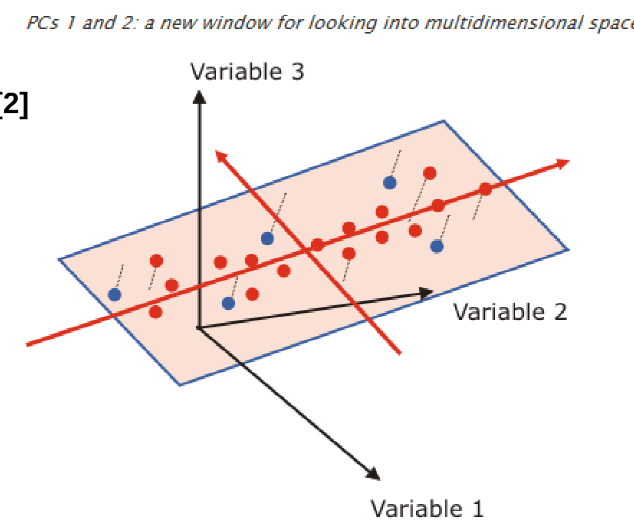
PCA basics – description of method

- Figures below: data with 3 variables and a number of objects
- Each row is represented as a point in three-dimensional coordinate system
- **Note**: not typical situation for use of PCA, since only 3 variables in data set.
    - However, appropriate for illustration
    - For matrices for more than 3 variables PCA cannot be visualised graphically, but mathematics are identical

[1]



[2]

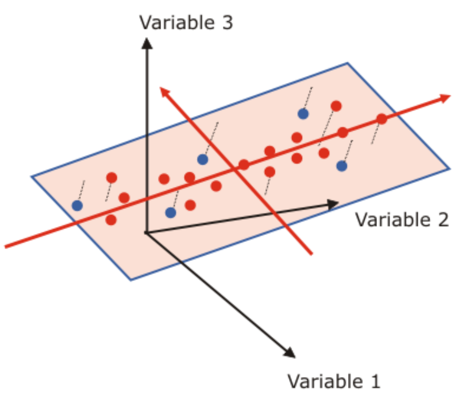


[1] & [2] Figures from CAMO Unscrambler software help documentation

PCA basics – description of method

- **Step 1**: compute averages of all variables (large point in plot below)
- **Step 2**: subtract averages from their corresponding variables. This is often called "*data centering*". This corresponds to moving the origin of coordinate system or vector space to average data point $\overline{x}$
- **Step 3**: search for direction in space that has **largest** variance ➔ component 1
- **Step 4**: search for direction in space that has **largest** variance **AND** is **orthogonal** to component 1 ➔ component 2
- **Step 5**: search for direction in space that has **largest** variance **AND** is **orthogonal** to component 1 **and** component 2 ➔ component 3
- **Step 6**: etc.

PCs 1 and 2: a new window for looking into multidimensional space **[1]**



- **Maximum** possible number of components
  - $min(K, N-1)$

[1] Figure from T. Næs, P.B. Brockhoff, O. Tomic, *Statistics for Sensory and Consumer Science*, (2010)

PCA basics

Data → Systematic variation + Residuals

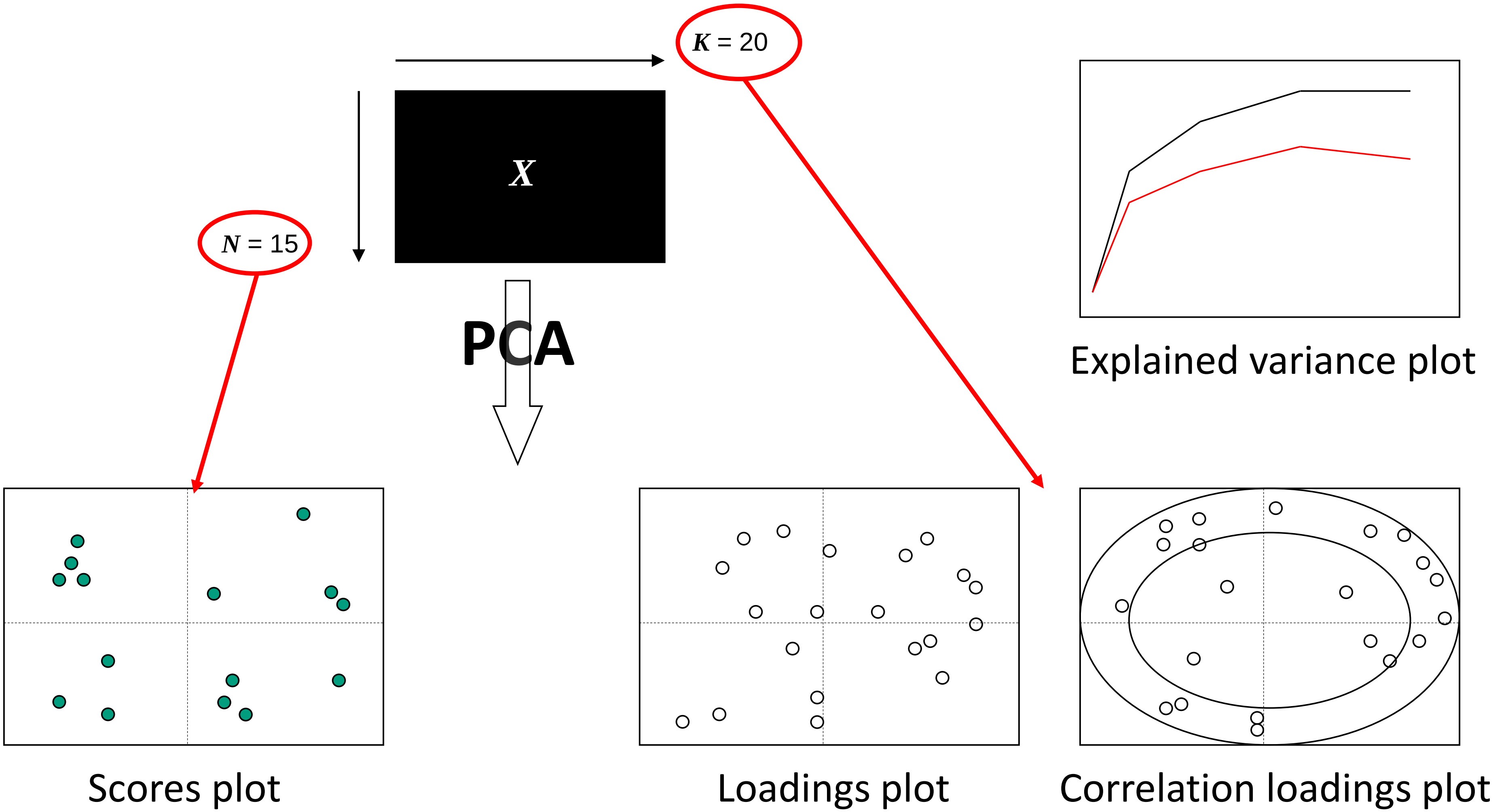
$$X = TP' + E$$

X: data matrix
T: PCA scores matrix
P: PCA loadings matrix
E: residuals / noise

Principal Components (PC's) describing the systematic variation in the data

PCA basics

Data in this illustration consists of 15 observations and 20 variables ➔ data matrix $\boldsymbol{X}$ of dimension (15×20)



PCA – scores and **loadings**

PCA basics – scores and loadings

- **Score plot**
  - is a scatter plot of columns of $\boldsymbol{T}$
  - objects close to each other have similar overall properties
  - objects far apart are very different
  - New coordinate system in a more compact subspace which spans the major variations

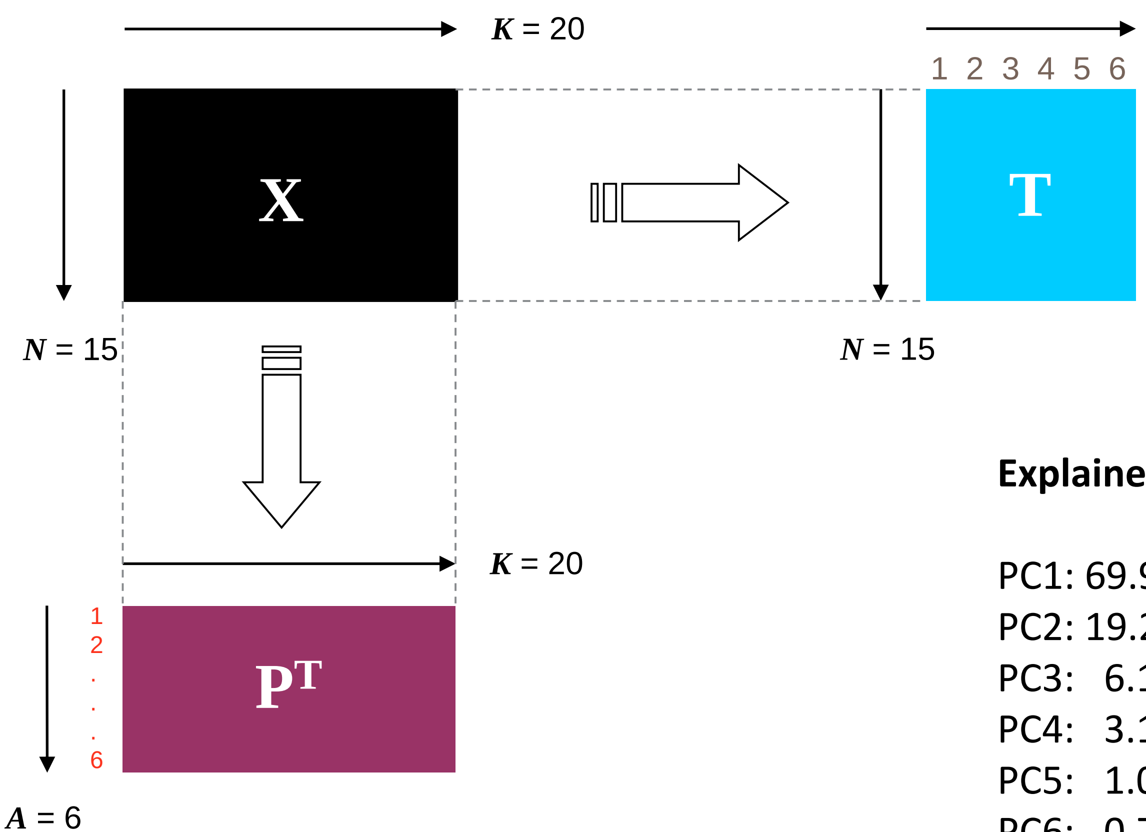
- **Loading plot**
  - is a scatter plot of rows of $\boldsymbol{P}^T$ (or columns of $\boldsymbol{P}$)
  - variables close to each other are highly correlated
  - Variables on opposite side of each other are highly negatively correlated

PCA basics – scores and loadings

- **Score plot** and **loading plot** together (so-called **bi-plot**)
  - Samples to the **right** in the scores plot are dominated by (have large values of) variables to the **right** in the loadings plot; objects at the **top** of the scores plot are dominated by variables at the **top** in the loadings plot; etc.

- Usually, **two-dimensional** plots with the **first two** components are used
- **Three-dimensional** plots of **first three** components are also possible (on screen with rotation)
- It is also possible to plot component 1 vs. component 2 in one plot and components 2 vs. component 3 in another plot, etc.

PCA basics – scores and loadings



T: PCA scores matrix

Explained variance in each component

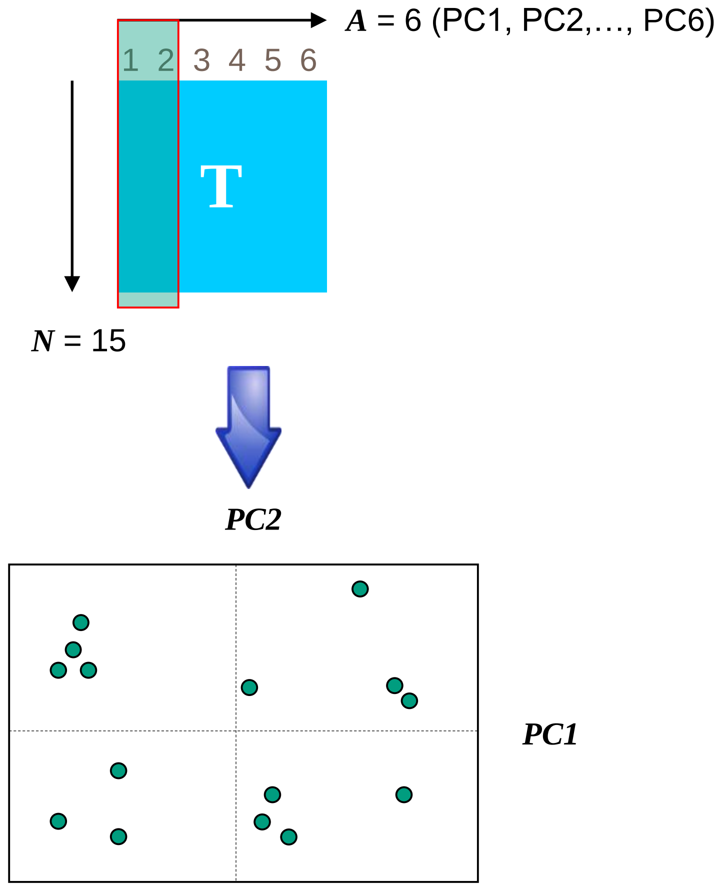
PC1: 69.9%
PC2: 19.2%
PC3: 6.1%
PC4: 3.1%
PC5: 1.0%
PC6: 0.7%

Total: 100%

P: PCA loadings matrix
A: number of components computed

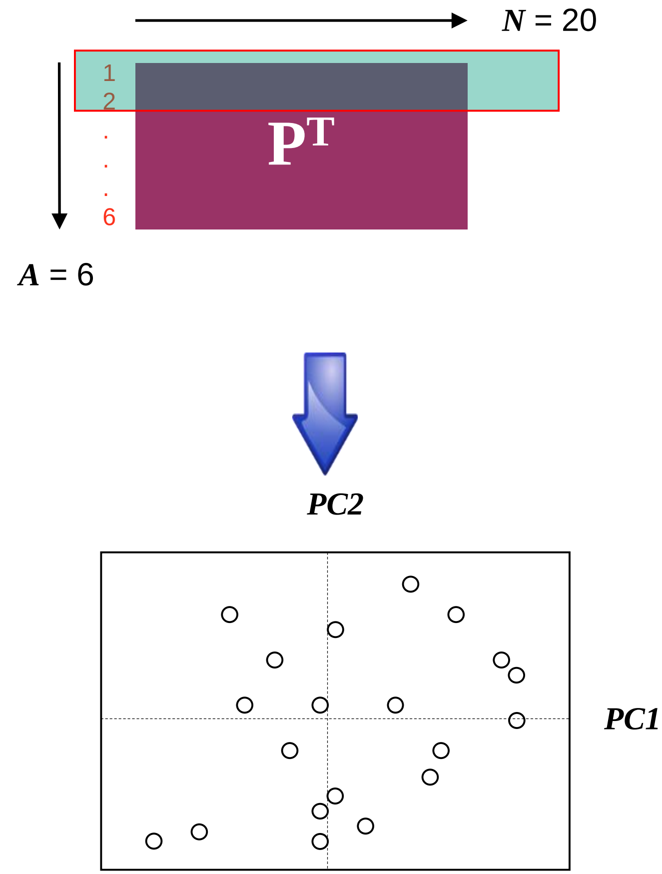
PCA basics – plotting scores and loadings

T: PCA scores matrix



Scores plot

P: PCA loadings matrix



Loadings plot

PCA – explained variance

PCA basics – explained variance

- Important information on **how much** variance (%) each component explains of the **total variance** in $\boldsymbol{X}$
- Explained variance by each component
    - **Highest** explained variance for **first** component
    - **Second**-highest explained variance for **second** component
    - **Third**-highest explained variance for **third** component
    - Etc.
- The higher the component, the higher the chance that it is unstable (since based on very small variances)

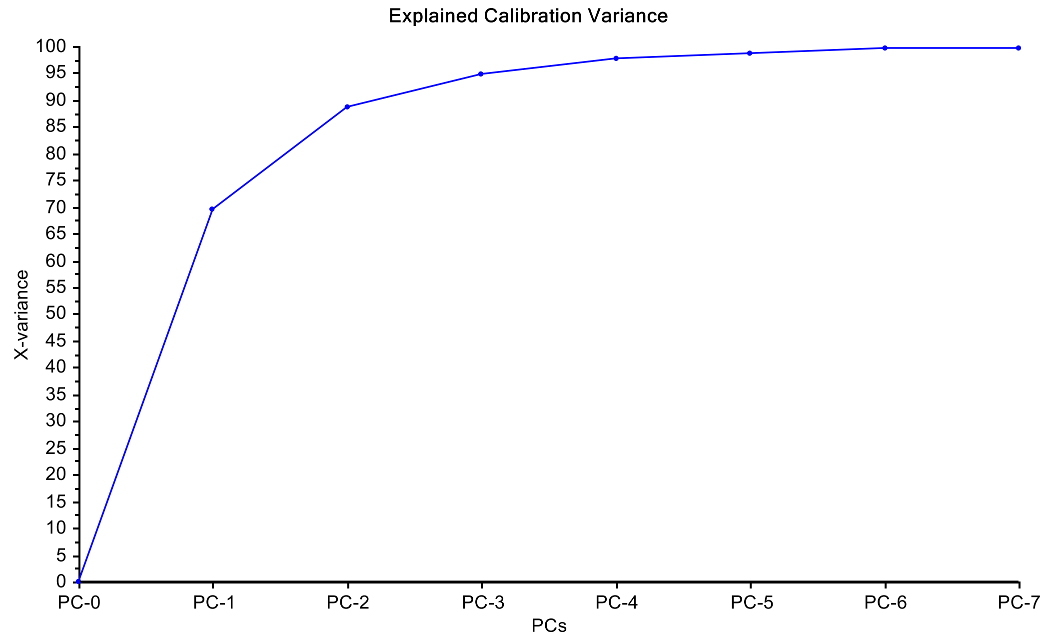
PCA basics – explained variance

Calibrated explained variance at each component

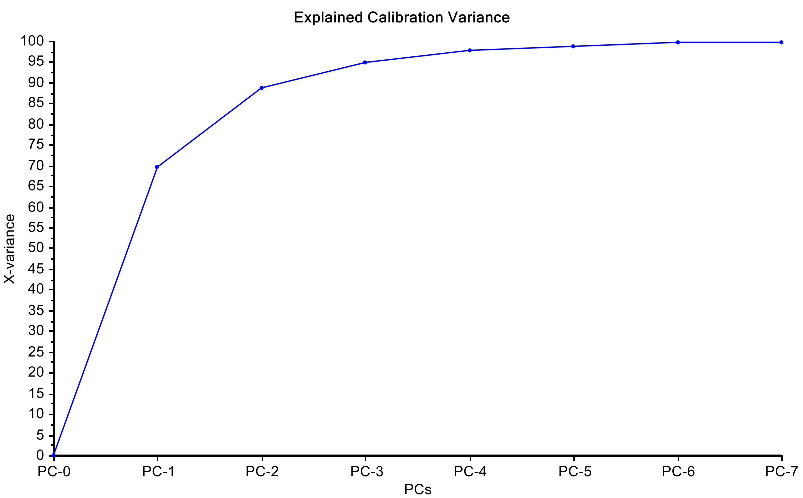
PC1: 69.9%
PC2: 19.2%
PC3: 6.1%
PC4: 3.1%
PC5: 1.0%
PC6: 0.7%

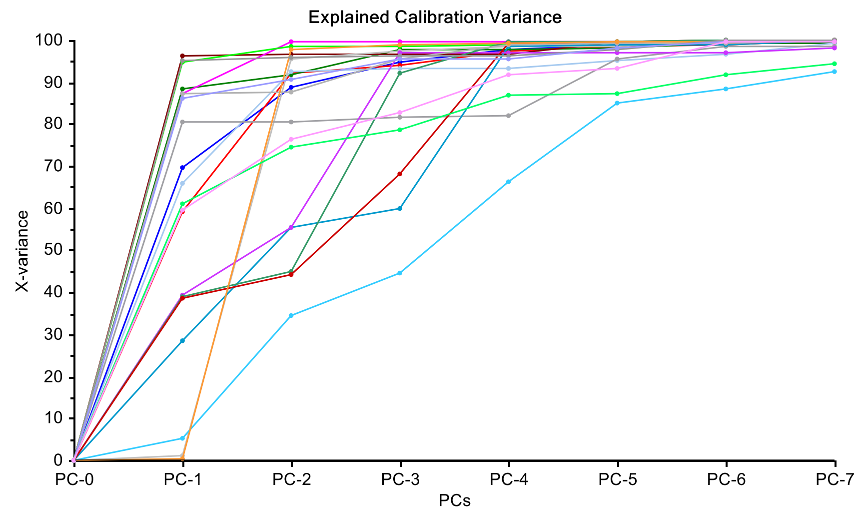
Total: 100%

Calibrated cumulative explained variance at each component



PCA basics – explained variance





Calibrated cumulative explained variance at each component across **all variables**

Calibrated cumulative explained variance for **each variable indivdually**

More on **validated** explained variance below in «PCA - validation»

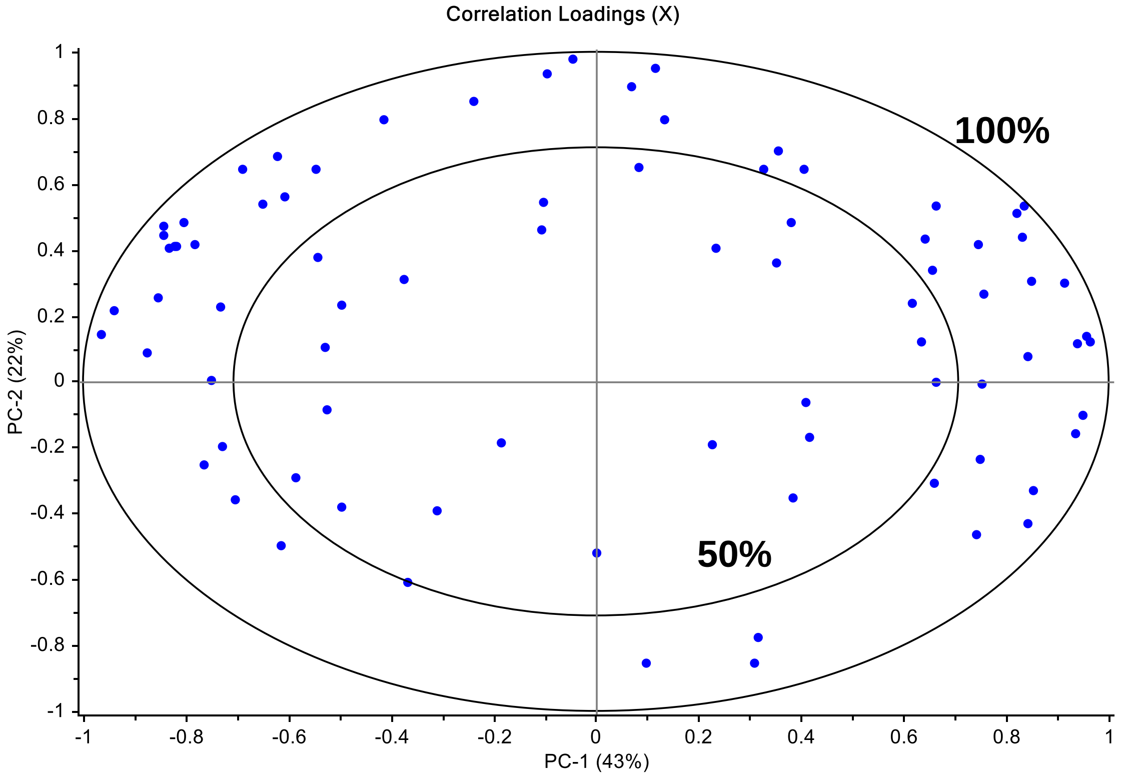
PCA – correlation loadings

PCA basics – correlation loadings

- Correlation loadings are a **modification** of the regular loadings ($\boldsymbol{P}$)
- Computed from **principal components** ($\boldsymbol{T}$) and **original variables** ($\boldsymbol{X}$)
- Graphical description of correlation loadings computation follows below
- **Advantage** of correlation loading plot
  - Provides direct information about on much the different variables are **correlated with** or **explained by** the different components
  - In particular, when the units of the variables are different, this may give additional and useful information
  - When variables are already **standardised**, the differences between the loadings and the correlation loadings plots will generally be **smaller**

PCA basics – correlation loadings

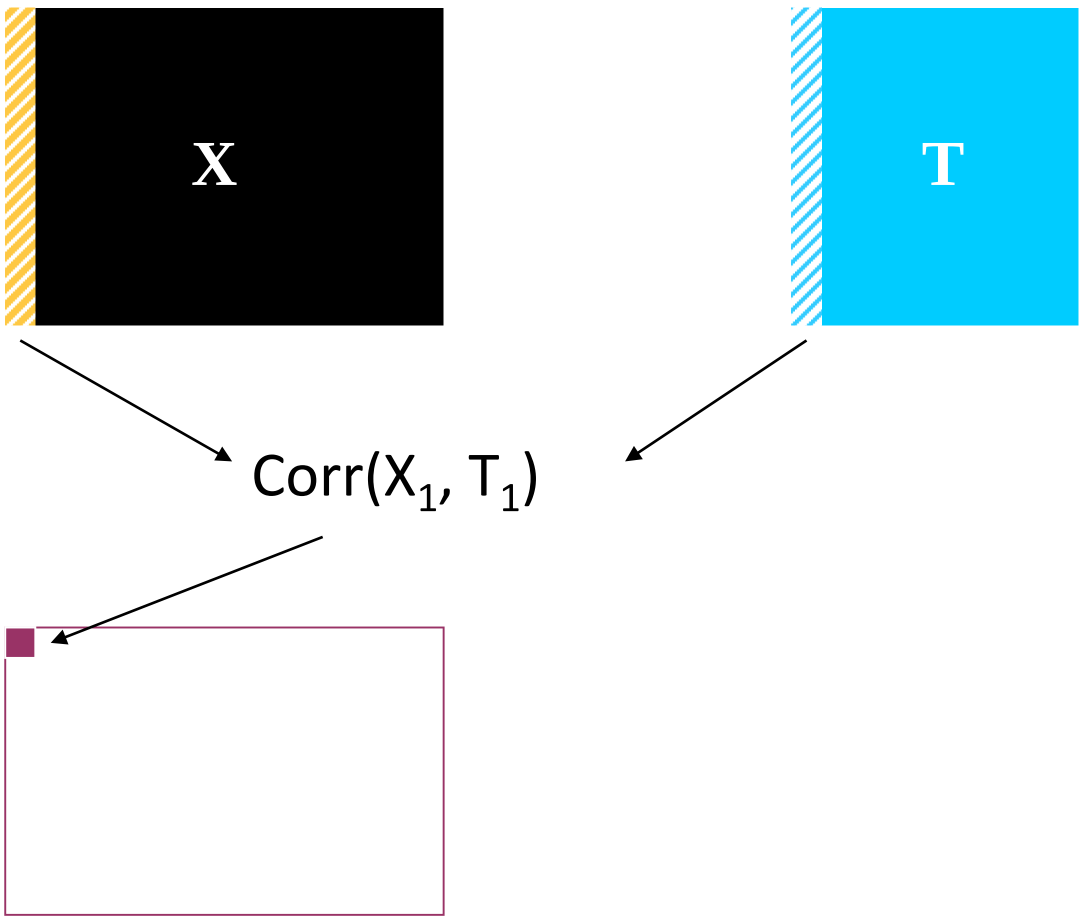
- Circles in the plot corresponding to various degrees of explained variances
- Typically one will present a circle for **100%** explained and for **50%** explained variance by the **two** components



PCA basics – computation of correlation loadings

X: Original data matrix

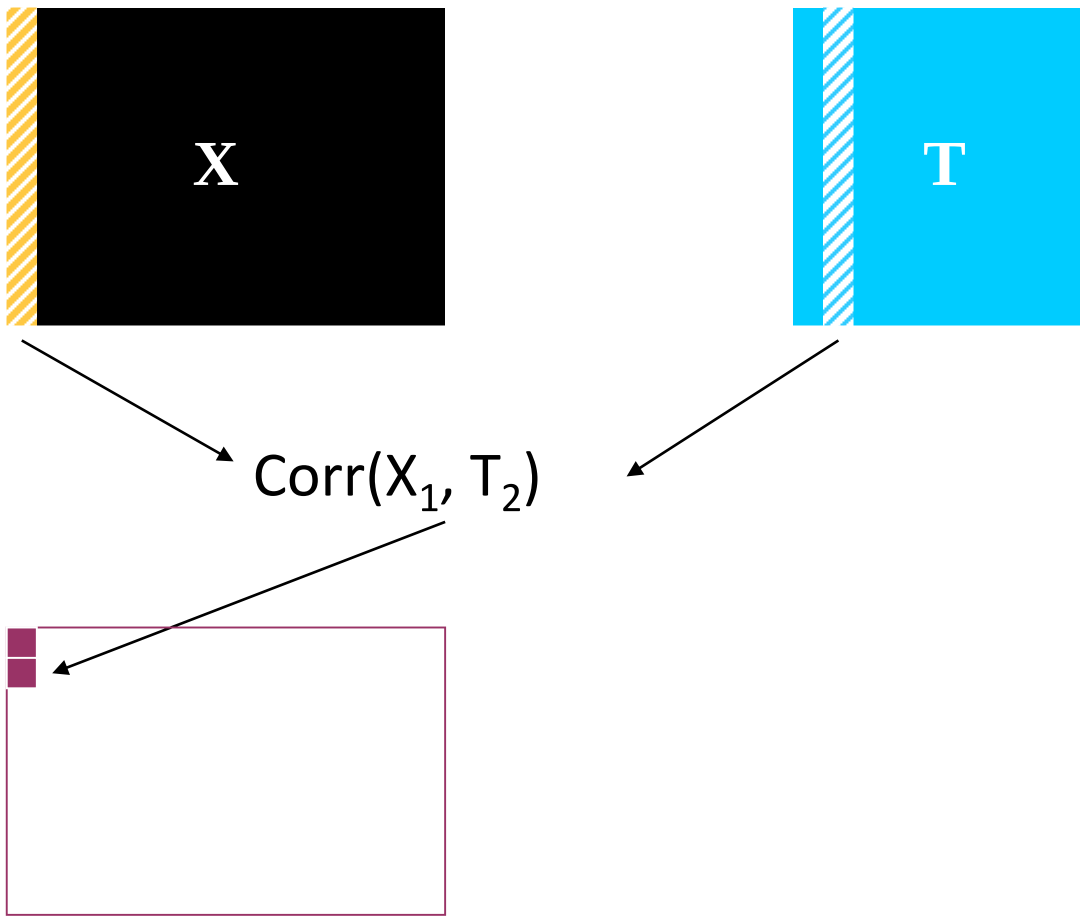
T: PCA scores matrix



PCA basics – computation of correlation loadings

X: Original data matrix

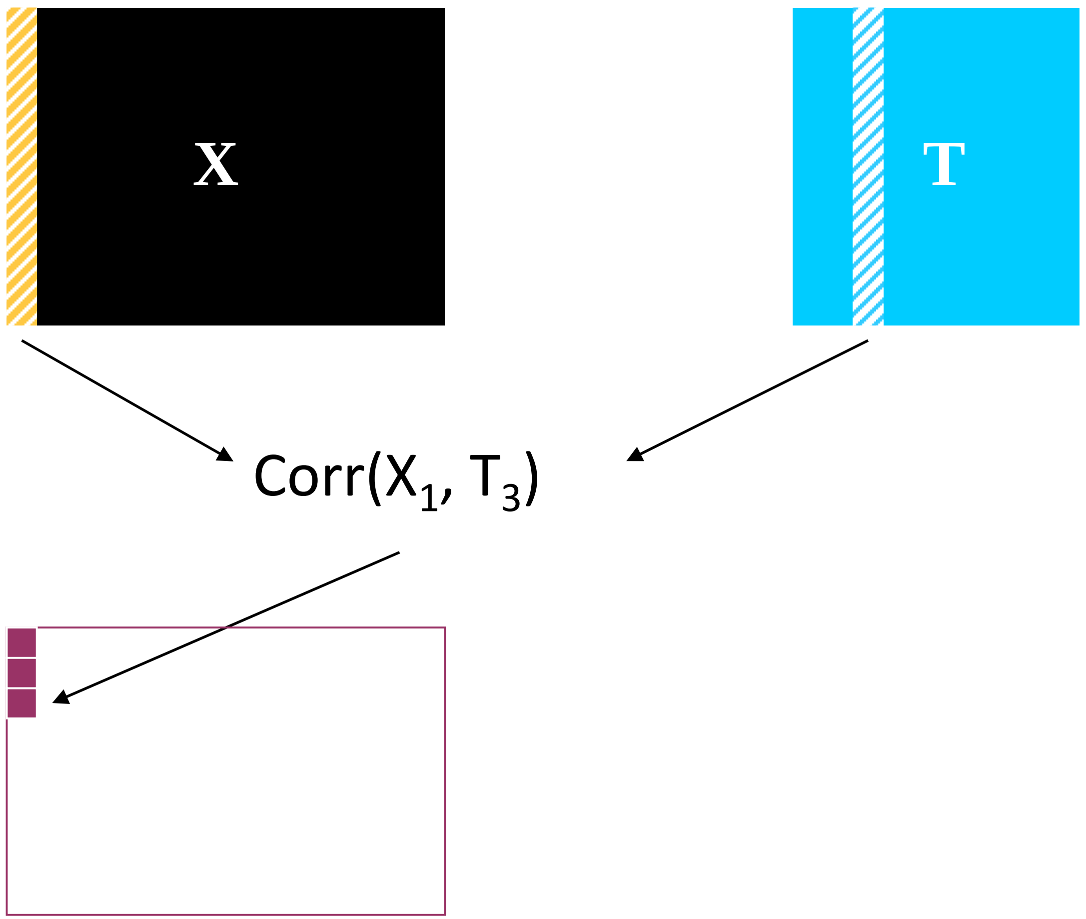
T: PCA scores matrix



PCA basics – computation of correlation loadings

X: Original data matrix

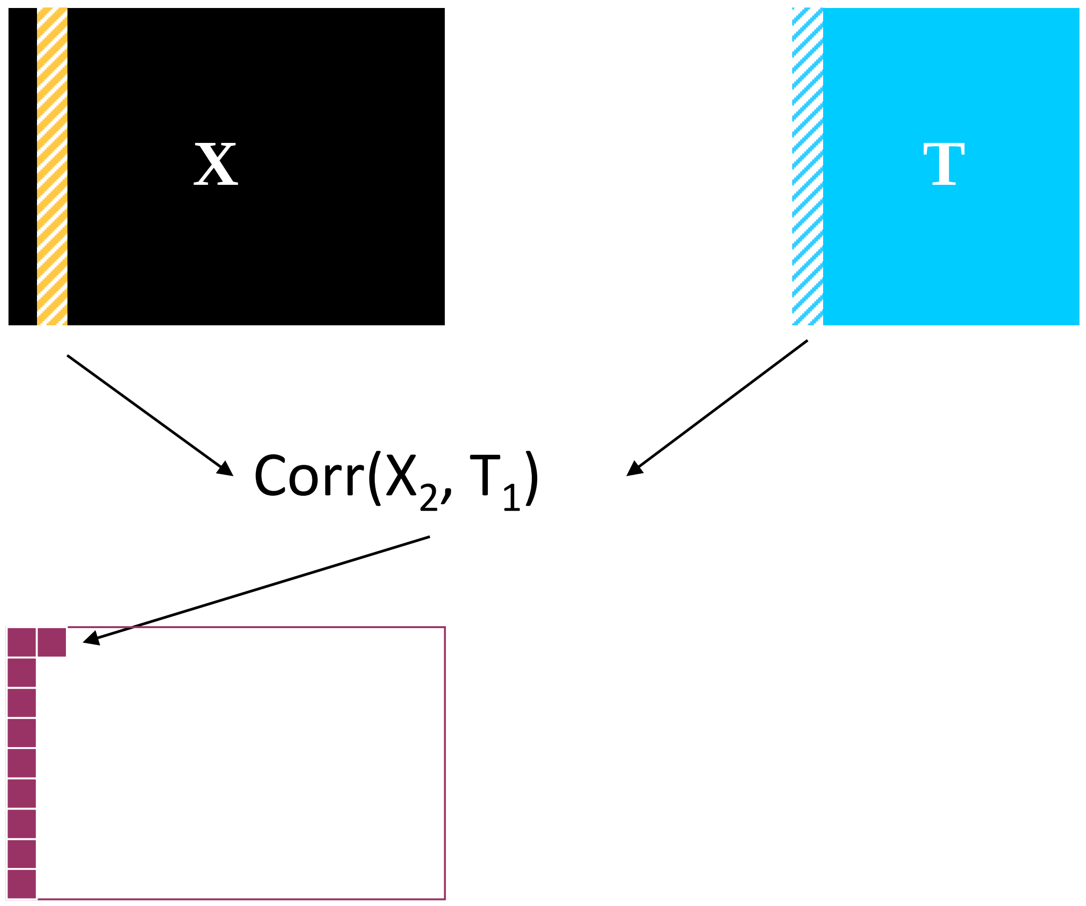
T: PCA scores matrix



PCA basics – computation of correlation loadings

X: Original data matrix

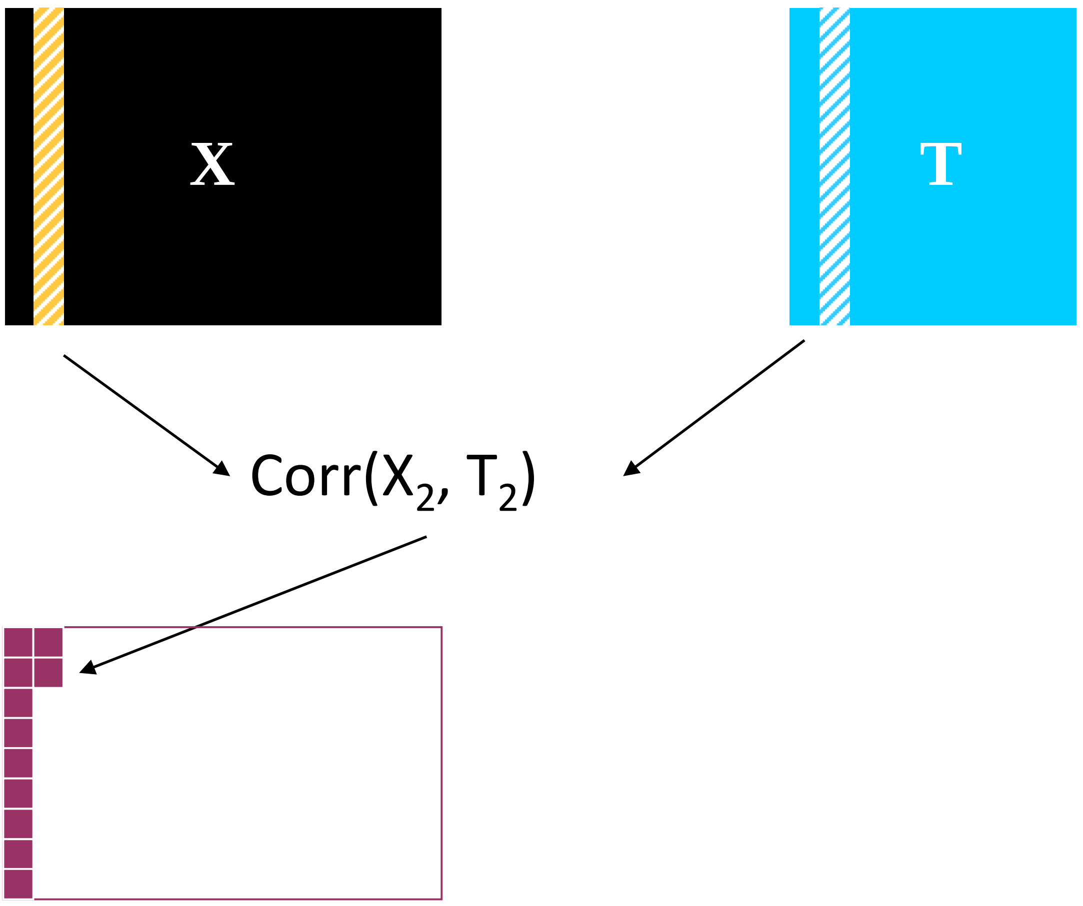
T: PCA scores matrix



PCA basics – computation of correlation loadings

X: Original data matrix

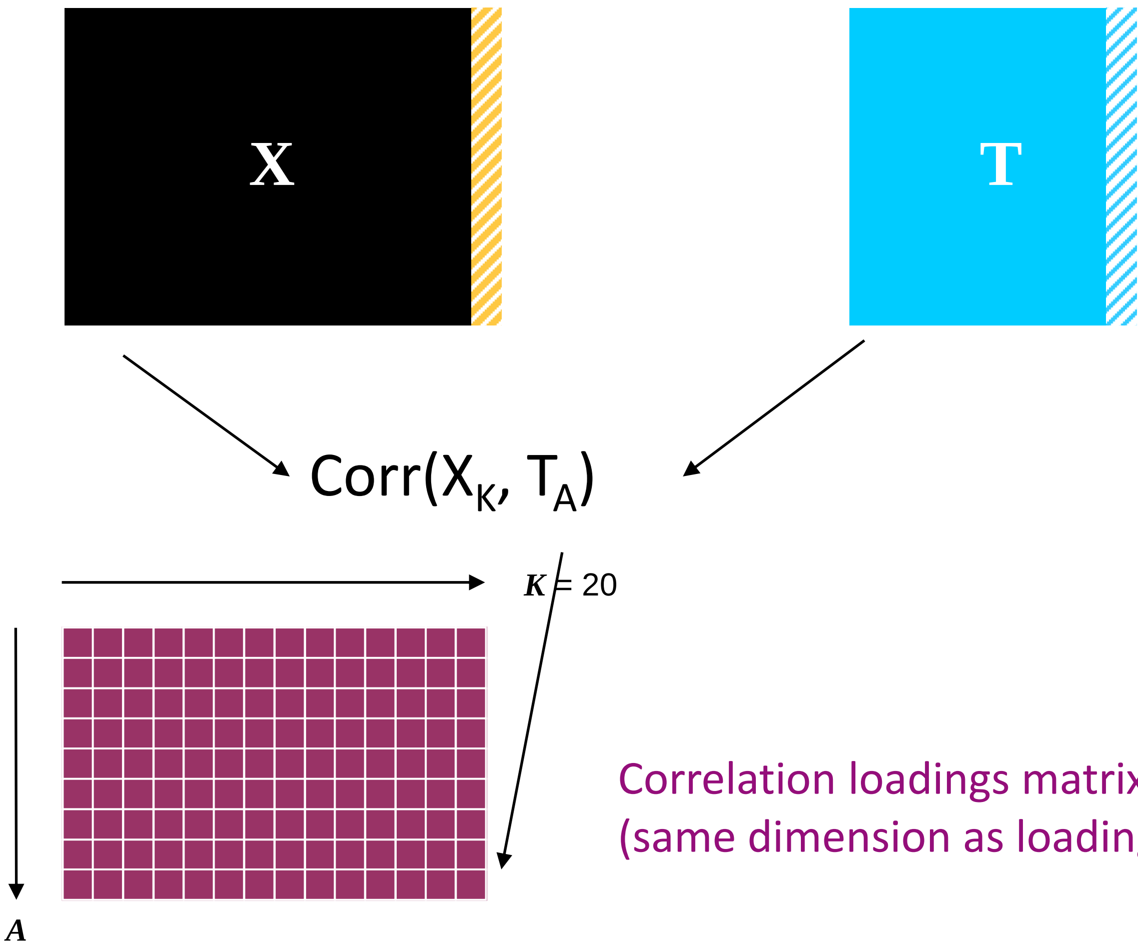
T: PCA scores matrix



PCA basics – computation of correlation loadings

X: Original data matrix

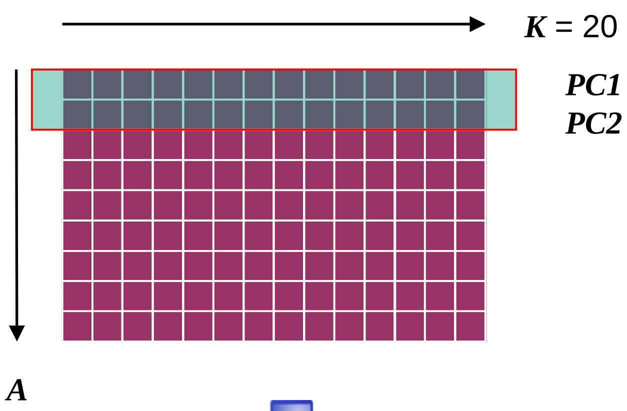
T: PCA scores matrix

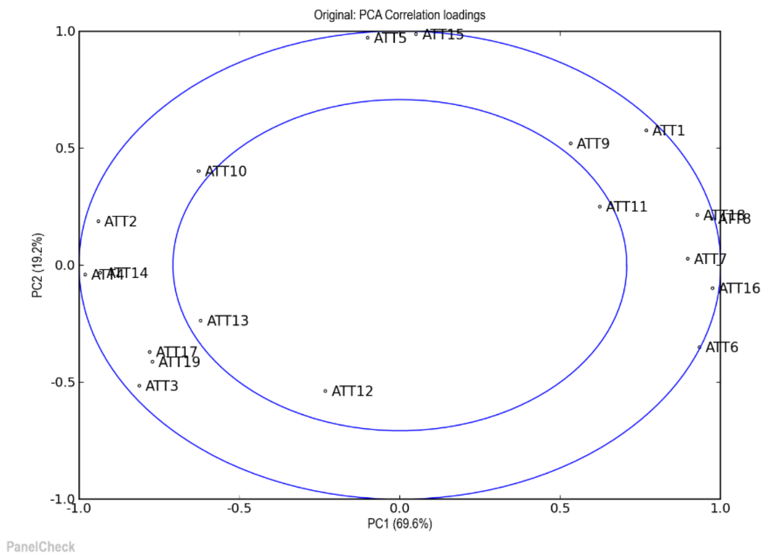


Correlation loadings matrix
(same dimension as loadings matrix ***P***)

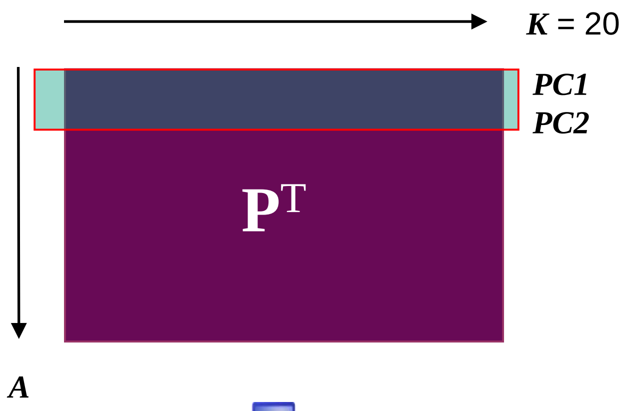
PCA basics – plotting correlation loadings

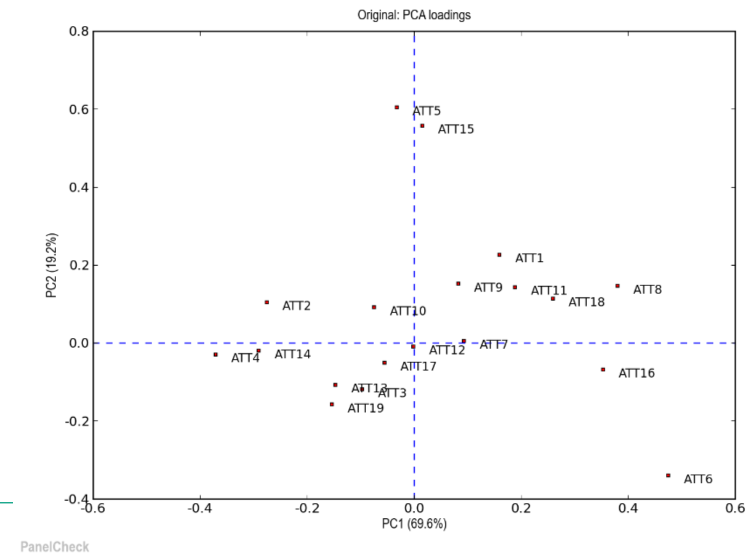
Correlation loadings matrix





Loadings matrix





PCA – centring and **standardisation**

PCA basics – centring and standardisation

- Only purpose of PCA is to look for directions with **high variance**
- This implies: if there are variables $\boldsymbol{x}_k$ in $\boldsymbol{X}$ that have a **larger variance** than others …
  - they will be given **most** attention
  - ➔ They will **dominate** the extracted components
  - ➔ They will **dominate** the plots
- Generally one is interested in letting all variables play a role in the estimation of components (there are exceptions) ➔ standardise variables $\boldsymbol{x}_k$ in $\boldsymbol{X}$
- Matrices in multivariate statistics are always **either** *centered* or *standardised*

PCA basics – centring and standardisation

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{NK} \end{pmatrix}$$

- Number of **objects (rows)**:
  - $n = 1 \ldots N$
- Number of **variables (columns)**:
  - $k = 1 \ldots K$
- Observed value x_{nk} for
  - n'th object
  - k'th variable

center

$$\boxed{x_{nk,cent} = x_{nk} - \bar{x}_k}$$

$$\bar{x}_k = \frac{1}{N}\sum_{n=1}^{N} x_{nk}$$

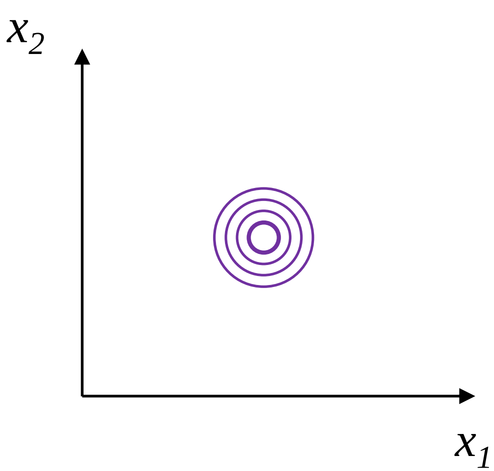
where

standardise

$$\boxed{x_{nk,stand} = \frac{x_{nk} - \bar{x}_k}{\sigma_k}}$$

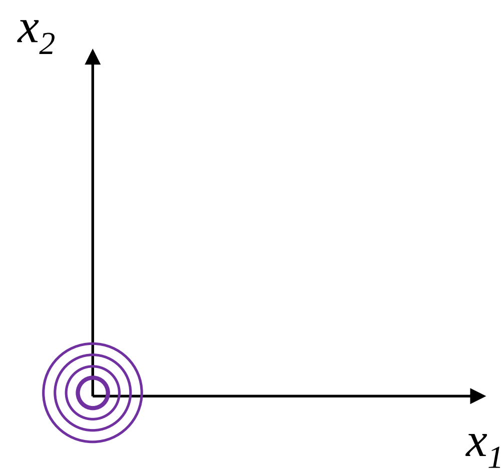
$$\sigma_k = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{nk} - \bar{x}_k)^2}$$

PCA basics – centring and standardisation



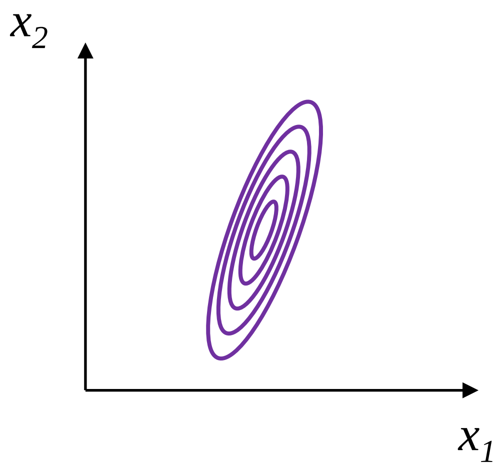
center data

$$x_{nk,cent} = x_{nk} - \bar{x}_k$$



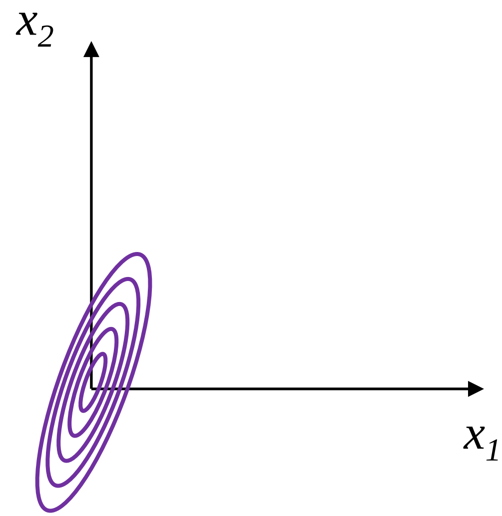
Equal variance of
x_1 and x_2

PCA basics – centring and standardisation



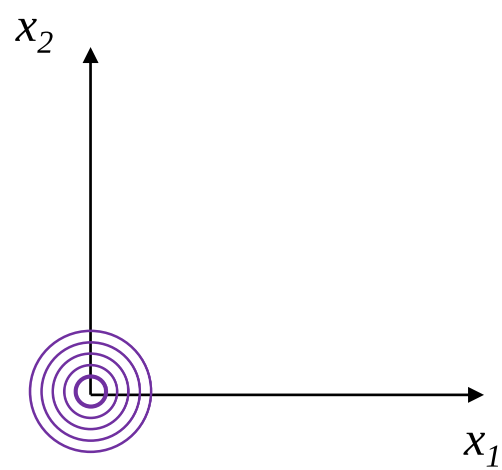
center data

$$x_{nk,cent} = x_{nk} - \bar{x}_k$$



standardise data

$$x_{nk,stand} = \frac{x_{nk} - \bar{x}_k}{\sigma_k}$$



Variance of x_2 is larger than variance of x_1

malahanobis

PCA basics – centring and standardisation

Person	Height (cm)	Weight (kg)	Shoe size
Person A	174	55	46
Person B	188	92	45
Person C	158	65	42
Person D	202	110	49
Person E	171	96	44
Person F	193	79	48
Mean	181	82.833333	45.6667
STD	16.198765	20.507722	2.58199

Original data

Person	Height (cm)	Weight (kg)	Shoe size
Person A	-7	-27.833333	0.33333
Person B	7	9.1666667	-0.66667
Person C	-23	-17.833333	-3.66667
Person D	21	27.166667	3.33333
Person E	-10	13.166667	-1.66667
Person F	12	-3.8333333	2.33333
Mean	0.00	0.00	0.00
STD	16.198765	20.507722	2.58199

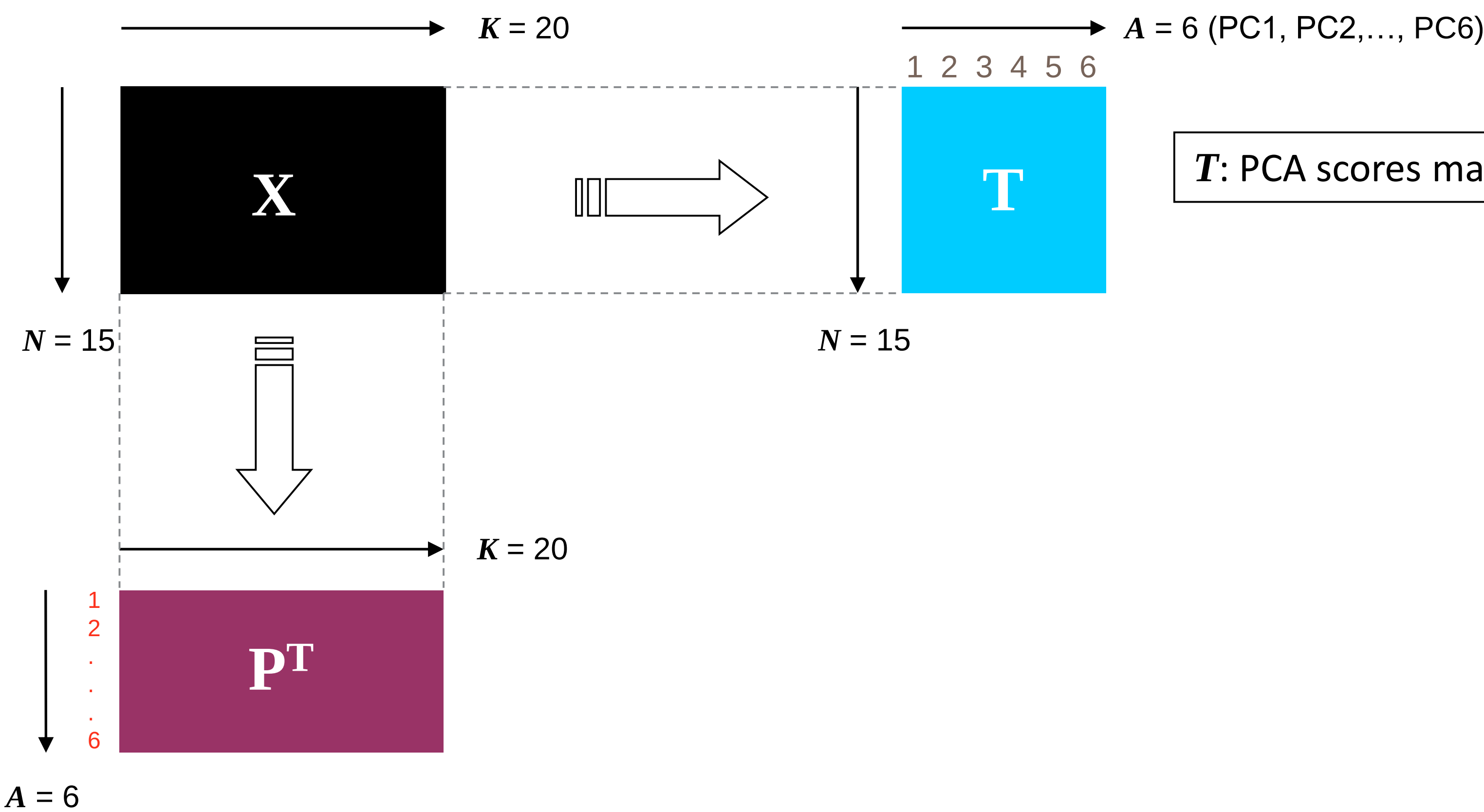
Centered data

Person	Height (cm)	Weight (kg)	Shoe size
Person A	-0.4321317	-1.3572123	0.1291
Person B	0.4321317	0.4469861	-0.2582
Person C	-1.4198613	-0.8695911	-1.42009
Person D	1.2963951	1.3247043	1.29099
Person E	-0.617331	0.6420346	-0.6455
Person F	0.7407972	-0.1869215	0.9037
Mean	0.00	0.00	0.00
STD	1	1	1

Standardised data

PCA – more on concept

PCA basics – more on concept



T: PCA scores matrix

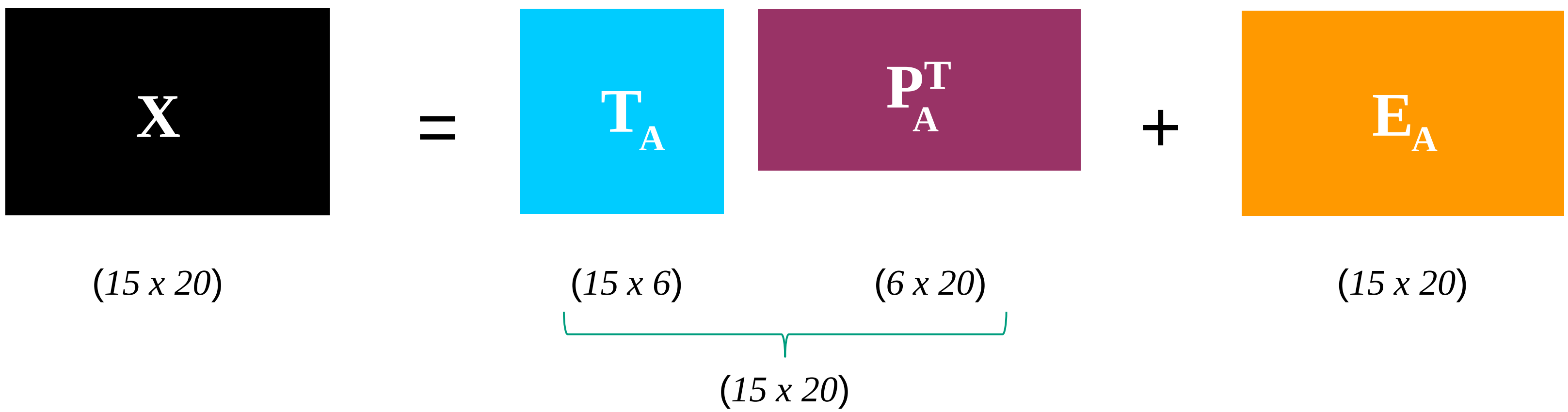
P: PCA loadings matrix
A: number of components computed

PCA basics

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$$

Example

$A = 6$

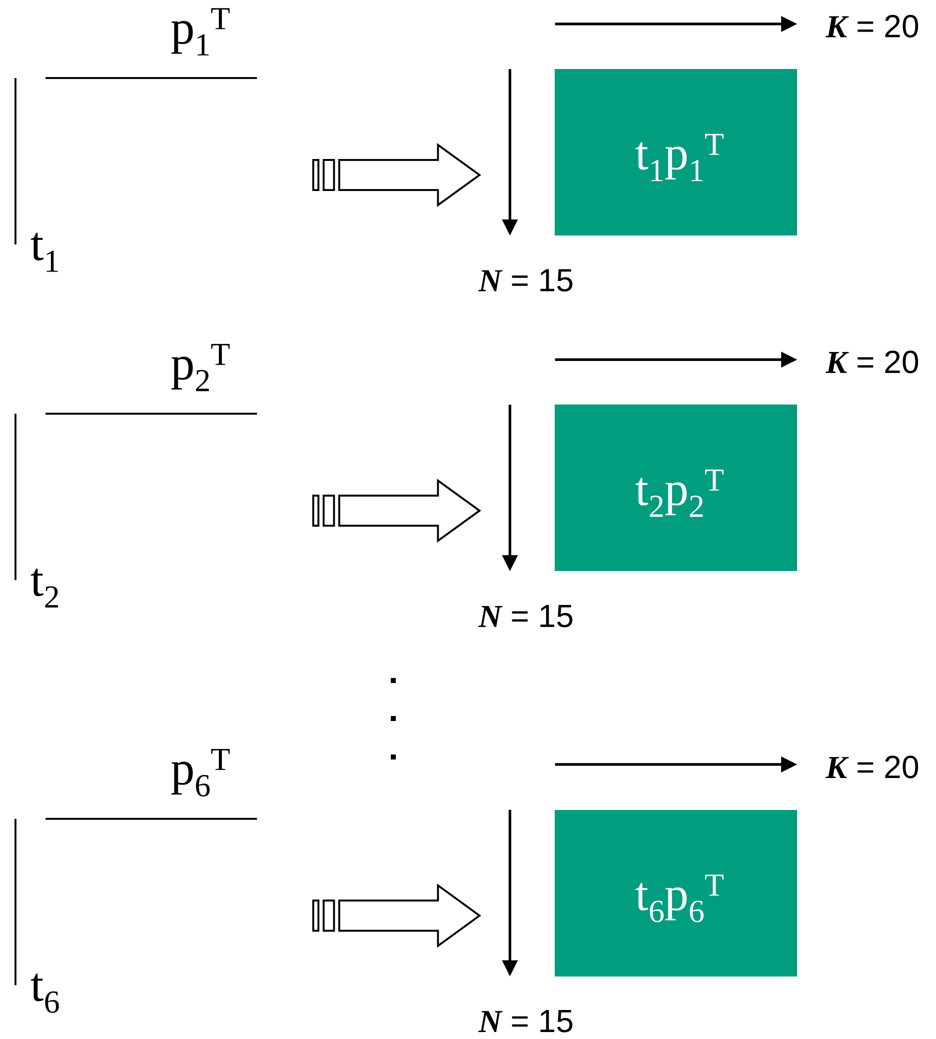


$$\boldsymbol{X} = \sum_{a=1}^{A} \boldsymbol{t}_a \boldsymbol{p}_a^T + \boldsymbol{E}$$

$$\boldsymbol{X} = \boldsymbol{t}_1 \boldsymbol{p}_1^T + \boldsymbol{t}_2 \boldsymbol{p}_2^T + \cdots + \boldsymbol{t}_A \boldsymbol{p}_A^T + \boldsymbol{E}$$

PCA basics – more on concept

From example: $X = t_1 p_1^T + t_2 p_2^T + \ldots + t_6 p_6^T$



Holds 69.9% of variance in X

Holds 19.2% of variance in X

Holds 0.7% of variance in X

PCA basics – more on concept

$$X = t_1p_1^T + t_2p_2^T + \ldots + t_6p_6^T$$

$$X = t_1p_1^T + t_2p_2^T + t_3p_3^T + t_4p_4^T + t_5p_5^T + t_6p_6^T$$

Explained variance in each component

PC1: 69.9%
PC2: 19.2%
PC3: 6.1%
PC4: 3.1%
PC5: 1.0%
PC6: 0.7%

Total: 100%

Systematic variation ?

Residuals ?

PCA basics

$$X = t_1p_1^T + t_2p_2^T + \ldots + t_6p_6^T$$

$$X = t_1p_1^T + t_2p_2^T + t_3p_3^T + t_4p_4^T + t_5p_5^T + t_6p_6^T$$

Explained variance in each component

PC1: 69.9%
PC2: 19.2%

PC3: 6.1%
PC4: 3.1%
PC5: 1.0%
PC6: 0.7%

Systematic variation ?

Residuals ?

Total: 100%

How many components are appropriate for the PCA model? **Validation!**

PCA basics

$$X = t_1 p_1^T + t_2 p_2^T + E$$

$$X = t_1 p_1' + t_2 p_2' + E$$

$$\hat{X} = t_1 p_1^T + t_2 p_2^T$$

$$\hat{X} = t_1 p_1^T + t_2 p_2^T$$

➔ $\hat{X}$ is a filtered, ”noise free” version of $\boldsymbol{X}$ (approximation of $\boldsymbol{X}$)

PCA – validation

PCA basics – validation

- Validation is necessary to gain knowledge on **how many** components are **appropriate** for the model, i.e. how many components can be used for …
  - Interpretation
  - Further analysis

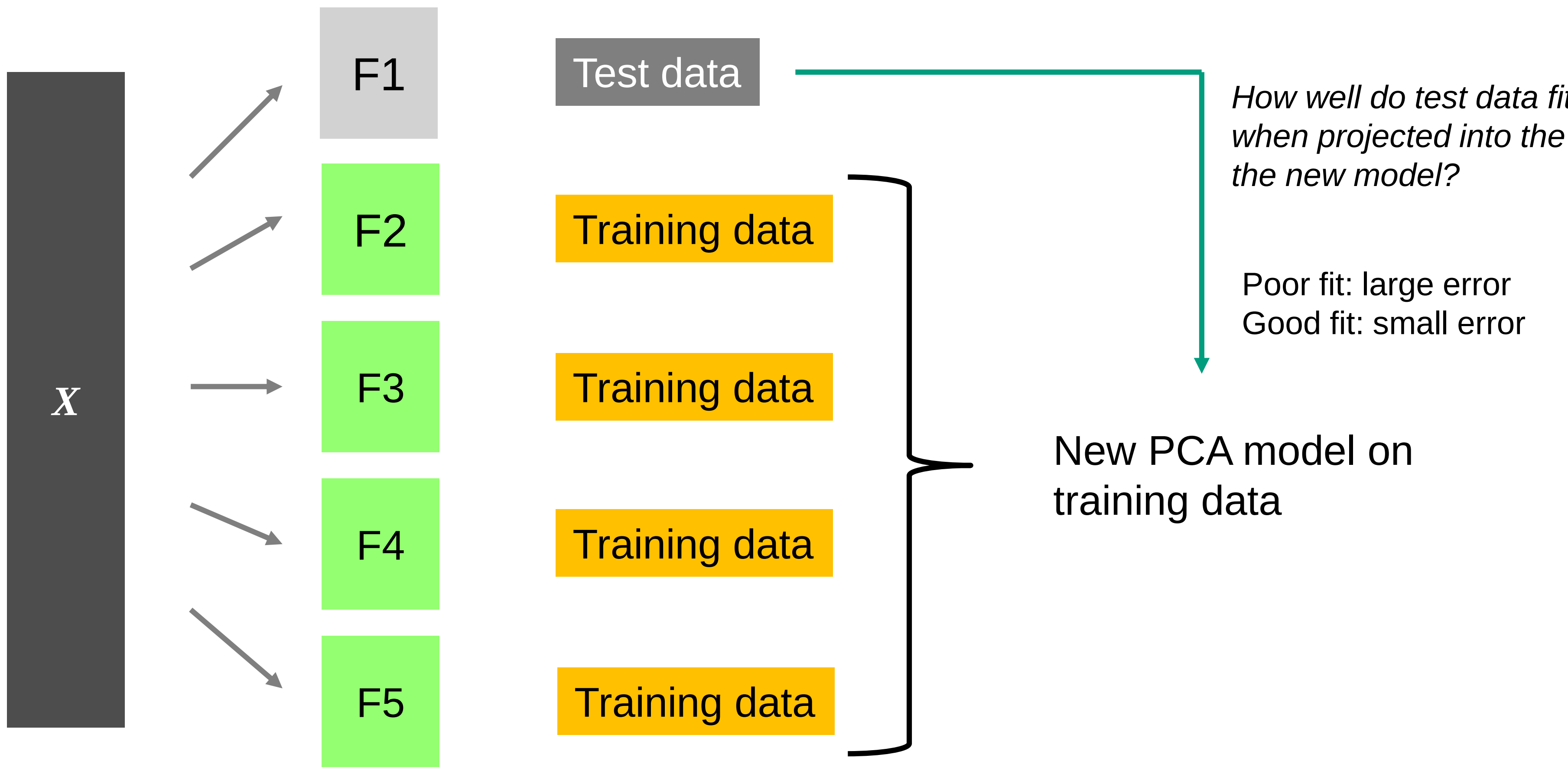
- Use of *internal cross validation* in PCA
  - **K-fold** cross validation (number of folds / splits used)
  - **LOO** cross validation (“Leave-one-out”)
    - LOO computationally more expensive compared to K-Fold
    - LOO is special case of K-Fold where K is equal to number of objects in data

PCA basics – validation

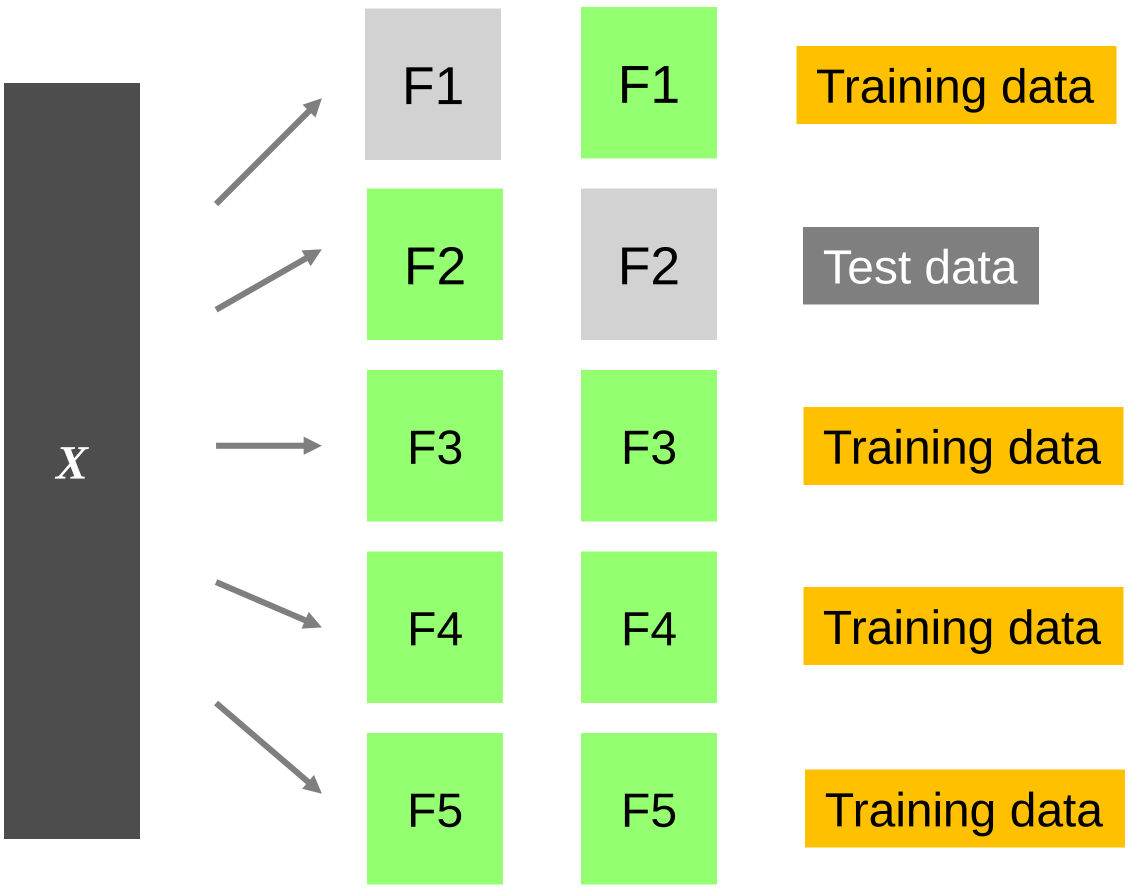
- More details on cross validation will be discussed in Ch. 6 – Learning Best Practices for Model Evaluation and Hyperparameter Tuning

- Use explained **validation** variance for choice of number of components
  - Point where curve of explained validated variance clearly flattens out ➔ point where one should stop interpreting components

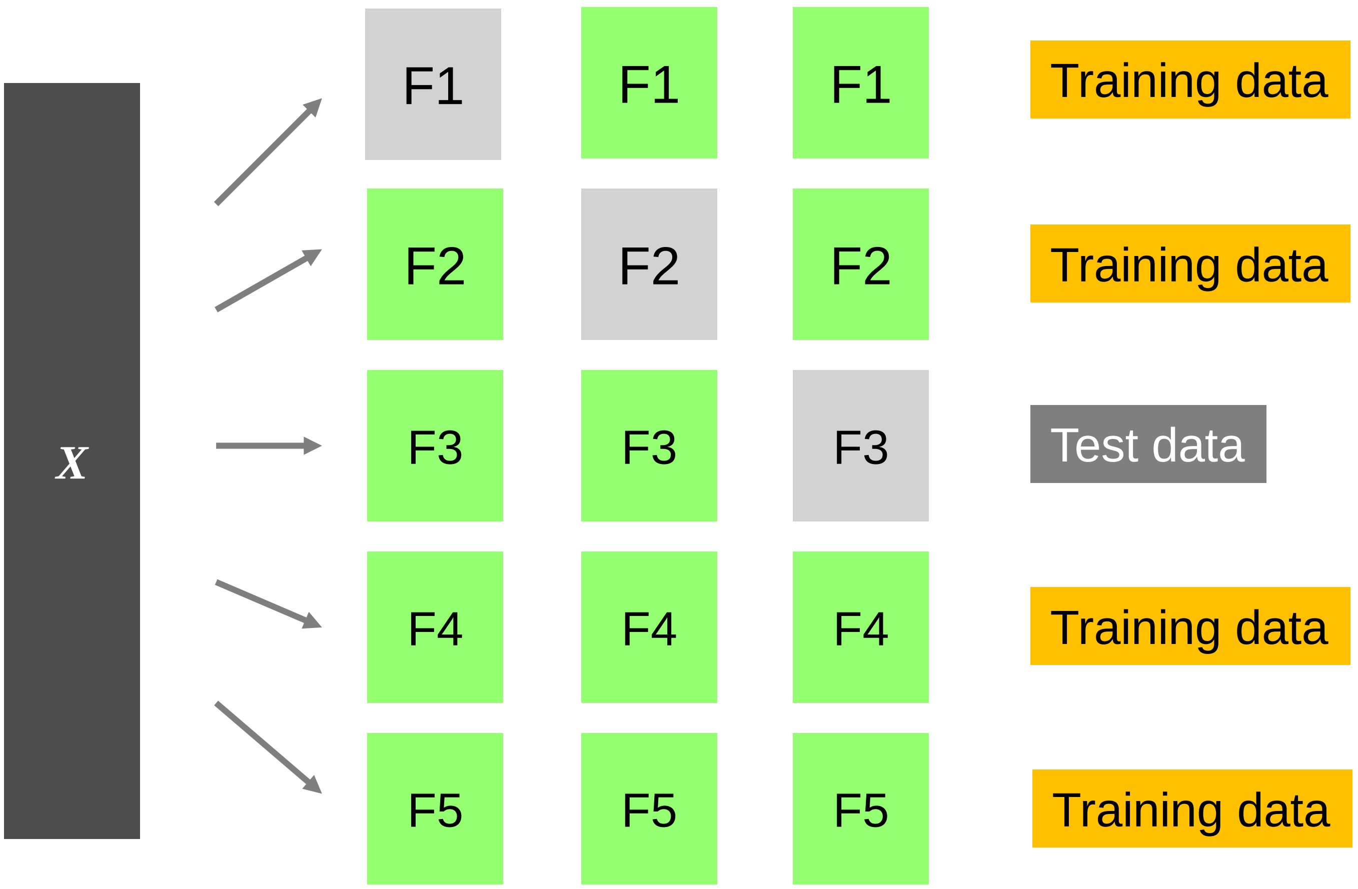
PCA basics - K-fold cross-validation process



PCA basics - K-fold cross-validation process



PCA basics - K-fold cross-validation process



PCA basics - K-fold cross-validation process

PCA basics - K-fold cross-validation process

K-fold cross-validation process

Error from **F1**

Error from **F2**

Error from **F3**

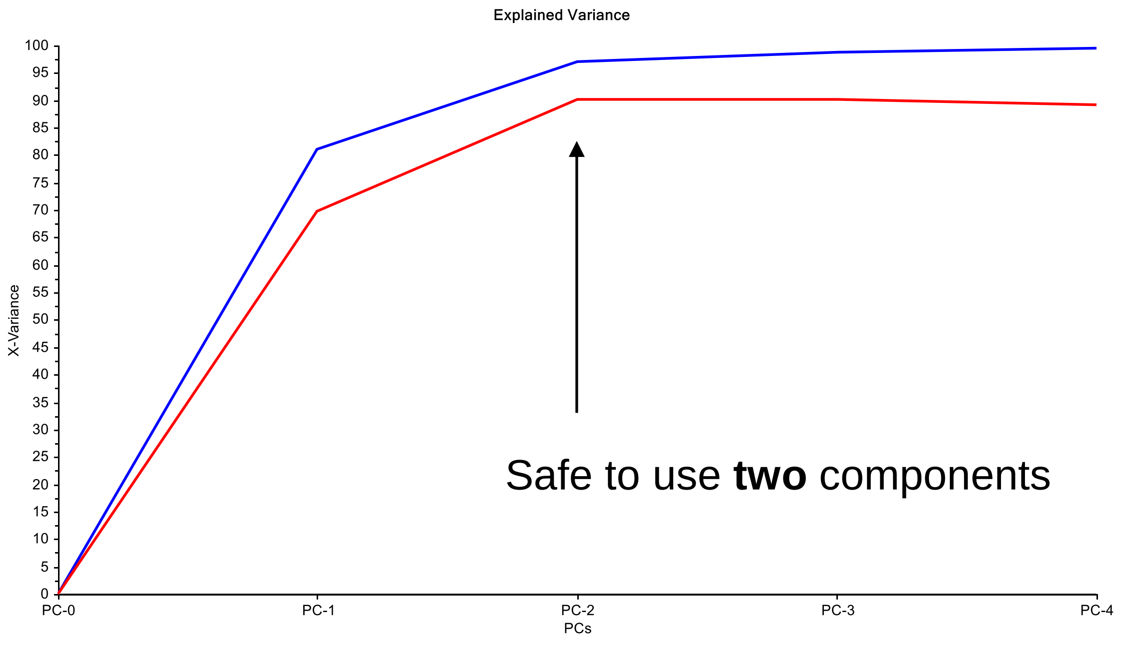
Error from **F4**

Error from **F5**

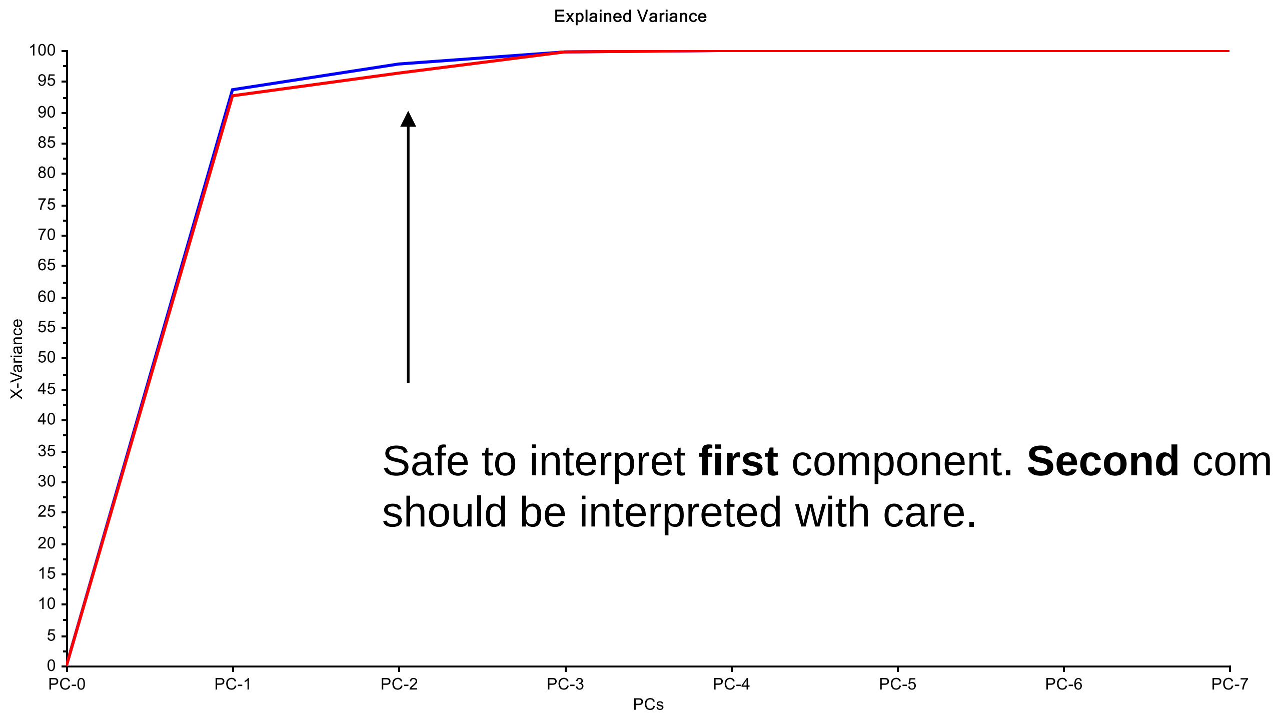
Compute average error across all folds

- Provides information on how many components are “safe” to use (explained validated variance plot)
- Provides information on the variance in the data

PCA basics – validation

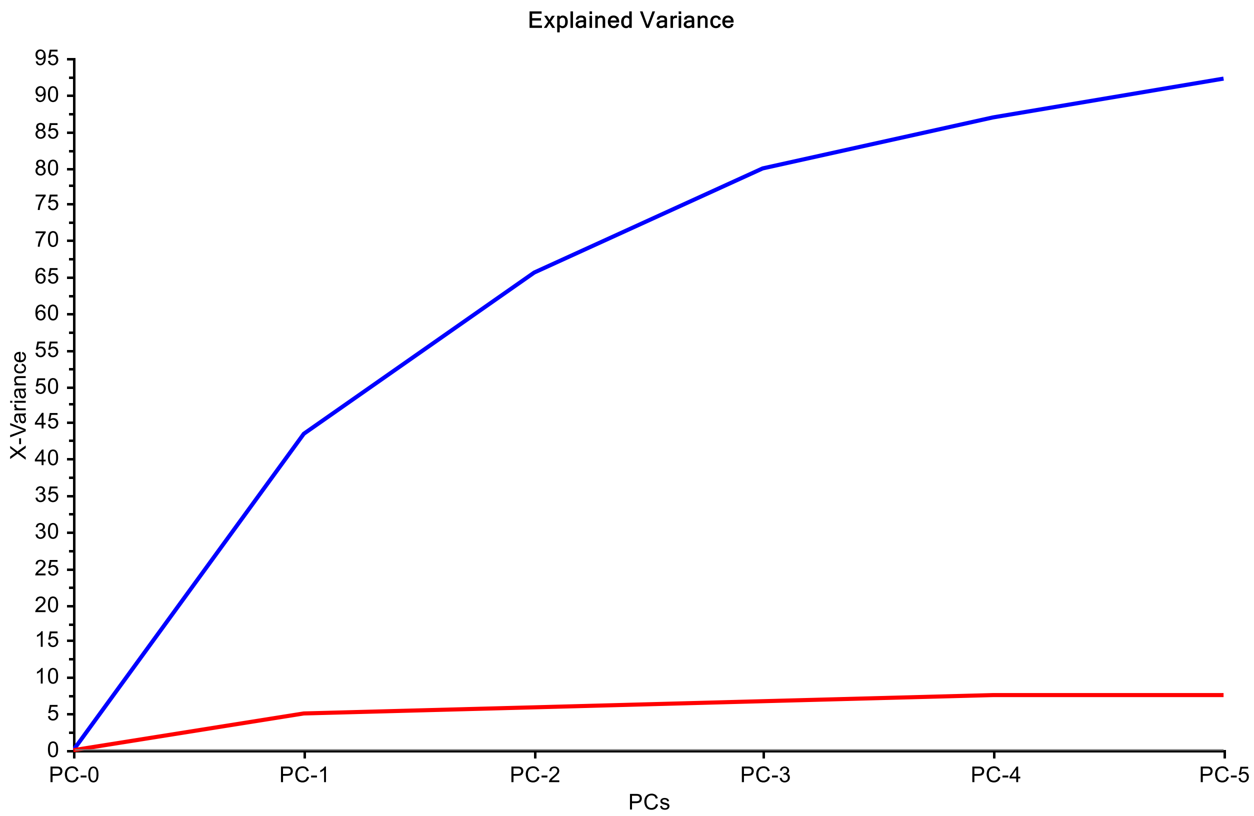


PCA basics – validation



Safe to interpret **first** component. **Second** component should be interpreted with care.

PCA basics – validation



Poor model – may be a result of few objects that are very different from each other or overfitting

PCA with **Hoggorm** and **HoggormPlot**

Hoggorm, HoggormPlot and examples

- **Hoggorm** package for multivariate statistics
  - GitHub: https://github.com/olivertomic/hoggorm
  - Read the Docs: http://hoggorm.readthedocs.io/en/latest/

- **HoggormPlot** package for convenient plotting of Hoggorm results
  - GitHub: https://github.com/olivertomic/hoggormPlot
  - Read the Docs: http://hoggormplot.readthedocs.io/en/latest/

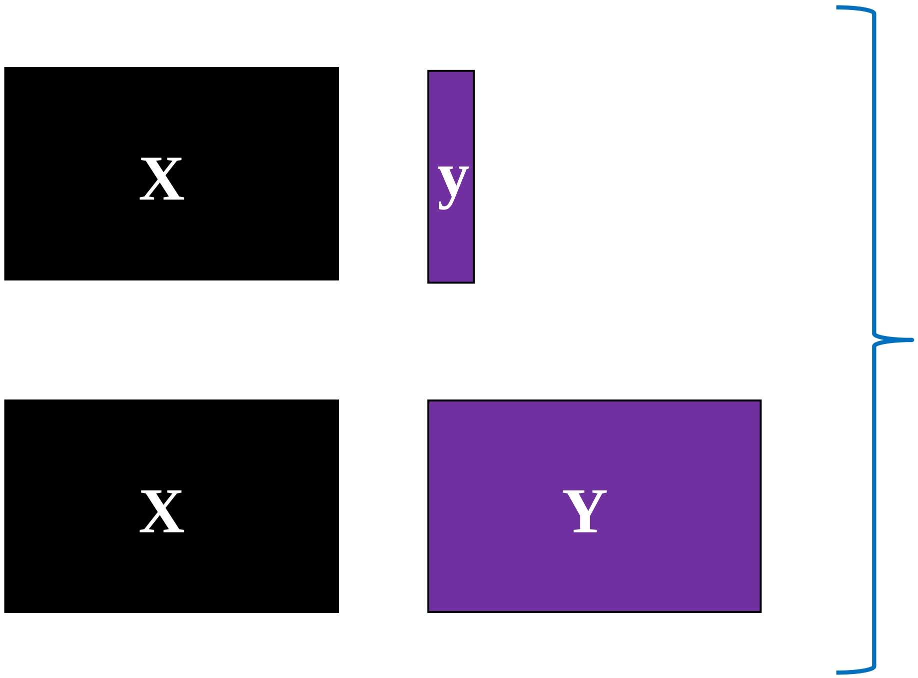
- **Examples** of how to use Hoggorm illustrated in Jupyter notebooks
  - GitHub: https://github.com/khliland/hoggormExamples

PRINCIPAL COMPONENT REGRESSION

Principal Component Regression (PCR)

- Analysis of **one** data table $\boldsymbol{X}$ (independent variables) and one **vector** $\boldsymbol{y}$ (response)
- Analysis of **two** data tables: $\boldsymbol{X}$ (independent variables) and $\boldsymbol{Y}$ (response)
- Idea behind PCR:
  - PCA on $\boldsymbol{X}$ followed by regression
  - Use first few components of $\boldsymbol{X}$ as base for regression analysis
  - All variability along the minor unstable principal component axes are thus disregarded in the regression analysis
- Solves the collinearity problem
- Is used for interpretation and prediction
- Provides tools for interpretation

PCR - basics



Regression
y or Y continuous variables

Classification
y contains classes

→ construct **dummy matrix**

PCR - basics

X and Y are assumed to be centred or standardised

$$\underset{(N\,x\,K)}{\mathbf{X}} = \underbrace{\underset{(N\,x\,A)}{\mathbf{T_A}}\;\underset{(A\,x\,K)}{\mathbf{P_A^T}}}_{(N\,x\,K)} + \underset{(N\,x\,K)}{\mathbf{E_A}}$$

$$\underset{(N\,x\,J)}{\mathbf{Y}} = \underbrace{\underset{(N\,x\,A)}{\mathbf{T_A}}\;\underset{(A\,x\,J)}{\mathbf{Q_A^T}}}_{(N\,x\,J)} + \underset{(N\,x\,J)}{\mathbf{F_A}}$$

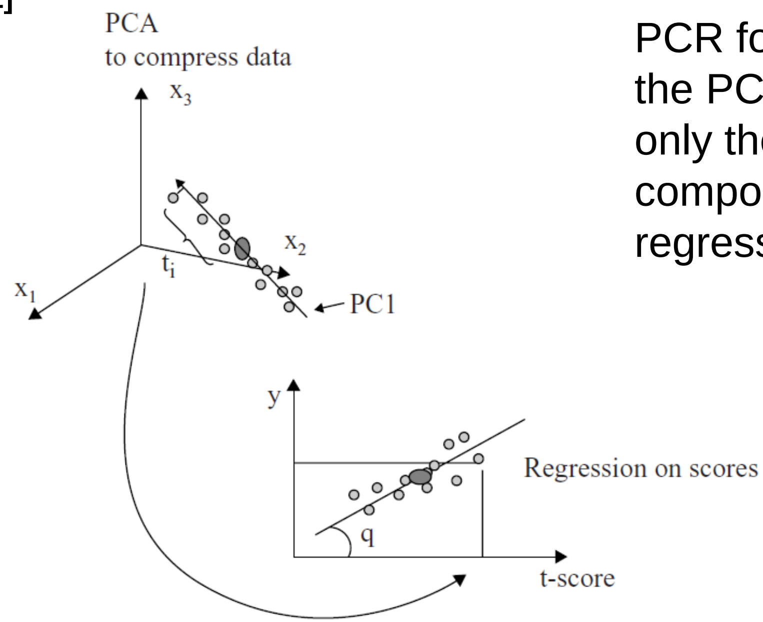
$\boldsymbol{X}$: independent variables
$\boldsymbol{T_A}$: scores
$\boldsymbol{P_A}$: X loadings
$\boldsymbol{E_A}$: X residuals

$\boldsymbol{Y}$: response matrix
$\boldsymbol{Q_A}$: Y loadings
$\boldsymbol{F_A}$: Y residuals

NOTE: scores $\boldsymbol{T}_A$ are acquired from PCA on $\boldsymbol{X}$

PCR - basics

[1]



PCR for data compression and regression. First, the PCA is used on X-data (upper left part) and only the information along the first few components (here only the first) is used for regression vs. the response y (lower right part).

[1] Figure 15.4 in T. Næs, P.B. Brockhoff, O.Tomic, *Statistics for sensory and consumer science* (2010)

PCR - basics

- Scores $\boldsymbol{T_A}$ are acquired by PCA on $\boldsymbol{X}$ $\boxed{\boldsymbol{X} = \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A}$
- Loadings $\boldsymbol{Q_A}$ in $\boxed{\boldsymbol{Y} = \boldsymbol{T}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A}$ are acquired by least squares method
  - $\boldsymbol{Q}_A^T = (\boldsymbol{T}_A^T \boldsymbol{T}_A)^{-1} + \boldsymbol{T}_A^T \boldsymbol{Y}$
- Predicting $\boldsymbol{Y}$ from new $\boldsymbol{X}$ using A components

$$\boxed{\widehat{\boldsymbol{Y}}_{new} = \boldsymbol{T}_{A,new} \boldsymbol{Q}_A^T = \boldsymbol{X}_{new} \boldsymbol{P}_A \boldsymbol{Q}_A^T}$$

$\boldsymbol{X}$: independent variables
$\boldsymbol{T}_A$: scores
$\boldsymbol{P_A}$: X loadings
$\boldsymbol{E_A}$: X residuals

$\boldsymbol{Y}$: response matrix
$\boldsymbol{Q_A}$: Y loadings
$\boldsymbol{F_A}$: Y residuals

NOTE: scores $\boldsymbol{T}_A$ are acquired from PCA on $\boldsymbol{X}$

PCR - basics

- Possible problem with PCR:
  - All components in model are extracted based on $\boldsymbol{X}$ only
  - This may be a drawback in situations where the first few components of $\boldsymbol{X}$ have less relation to $\boldsymbol{Y}$ than the components with minor variability

- A possible improvement over PCR: Partial Least Squares Regression (PLSR)

PARTIAL LEAST SQUARES REGRESSION

Partial Least Squares Regression (PLSR)

- Analysis of **one** data table $\boldsymbol{X}$ (independent variables) and one **vector** $\boldsymbol{y}$ (response)
  - PLS1 method
- Analysis of **two** data tables: $\boldsymbol{X}$ (independent variables) and $\boldsymbol{Y}$ (response)
  - PLS2 method
- Based on the same general model structure as PCR, however components are computed from $\boldsymbol{X}$ and $\boldsymbol{Y}$
- PLSR and PCR are used the same way from a practical point of view
- Solves the collinearity problem
- Is used for interpretation and prediction
- Provides tools for interpretation (same as for PCR)

PLSR - basics

- Obtaining PLS components
  - **Maximise covariance** between linear functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ (both centred as for PCR)
  - Effect of first factor is subtracted from $\boldsymbol{X}$ and $\boldsymbol{Y}$ → residuals are used for computing the second component
  - Procedure continues until the desired number of components, $\boldsymbol{A}$, has been extracted
- PLS components are orthogonal
- Components extracted in this way are more relevant for the prediction of $\boldsymbol{Y}$ than components found by PCR
- This may sometimes lead to models with a **smaller** number of components, which may possibly be easier to interpret

PLSR - basics

- The covariance criterion for PLSR is a **compromise** between the variance criterion used for PCR and the correlation criterion used for MLR
- Therefore, PLS is a **compromise** between the very stable and conservative PCR and the MLR which uses the $\boldsymbol{Y}$-information as actively as possible
- Note that the PLS solution for several $\boldsymbol{Y}$-values is not obtained by separate fitting of each individual $\boldsymbol{Y}$-variable

PLSR - basics

$$\underset{(N\,x\,K)}{\mathbf{X}} = \underbrace{\underset{(N\,x\,A)}{\mathbf{T_A}}\;\underset{(A\,x\,K)}{\mathbf{P_A^T}}}_{(N\,x\,K)} + \underset{(N\,x\,K)}{\mathbf{E_A}}$$

$$\underset{(N\,x\,J)}{\mathbf{Y}} = \underbrace{\underset{(N\,x\,A)}{\mathbf{T_A}}\;\underset{(A\,x\,J)}{\mathbf{Q_A^T}}}_{(N\,x\,J)} + \underset{(N\,x\,J)}{\mathbf{F_A}}$$

$\boldsymbol{X}$: independent variables
$\boldsymbol{T_A}$: scores
$\boldsymbol{P_A}$: X loadings
$\boldsymbol{E_A}$: X residuals

$\boldsymbol{Y}$: response matrix
$\boldsymbol{Q_A}$: Y loadings
$\boldsymbol{F_A}$: Y residuals

NOTE: scores $\boldsymbol{T_A}$ are acquired by maximising covariance between linear functions of $\boldsymbol{X}$ and linear functions of $\boldsymbol{Y}$

PLSR - basics

- $\boxed{\boldsymbol{X} = \boldsymbol{T}_A \boldsymbol{P}_A^T + \boldsymbol{E}_A}$
- $\boxed{\boldsymbol{Y} = \boldsymbol{T}_A \boldsymbol{Q}_A^T + \boldsymbol{F}_A}$

- Scores $\boldsymbol{T_A}$ are acquired maximising covariance between between linear functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$

- Predicting $\boldsymbol{Y}$ from new $\boldsymbol{X}$ using A components

$$\widehat{\boldsymbol{Y}}_{new} = \boldsymbol{X}_{new} \boldsymbol{B}_A$$

`regressionCoefficients(numComp=1)`

Returns regression coefficients from the fitted model using all available samples and a chosen number of components.

$\boldsymbol{X}$: independent variables
$\boldsymbol{T}_A$: scores
$\boldsymbol{P_A}$: X loadings
$\boldsymbol{E_A}$: X residuals

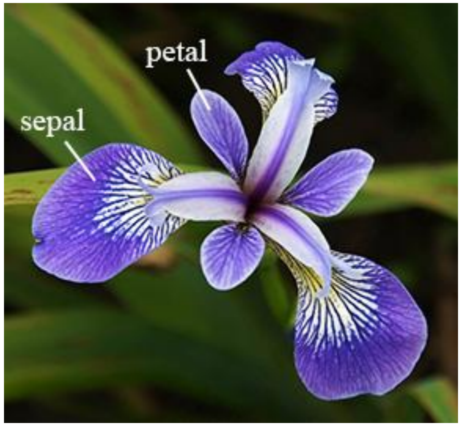
$\boldsymbol{Y}$: response matrix
$\boldsymbol{Q_A}$: Y loadings
$\boldsymbol{F_A}$: Y residuals

NOTE: scores $\boldsymbol{T_A}$ are acquired by maximising covariance between linear functions of $\boldsymbol{X}$ and linear functions of $\boldsymbol{Y}$

IRIS DATA - classification

The data used in examples

- Iris data set
  - Ronald A. Fisher
  - Collected in 1936



- Often used for classification / pattern recognition tutorials
  - Few features (4)
  - Few classes (3)
  - Simple domain
- https://archive.ics.uci.edu/ml/datasets/Iris

UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Iris data

Iris Setosa

50 instances

Iris Versicolor

50 instances

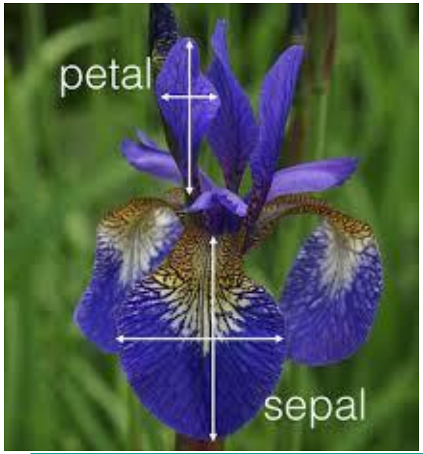
Iris Virginica

50 instances

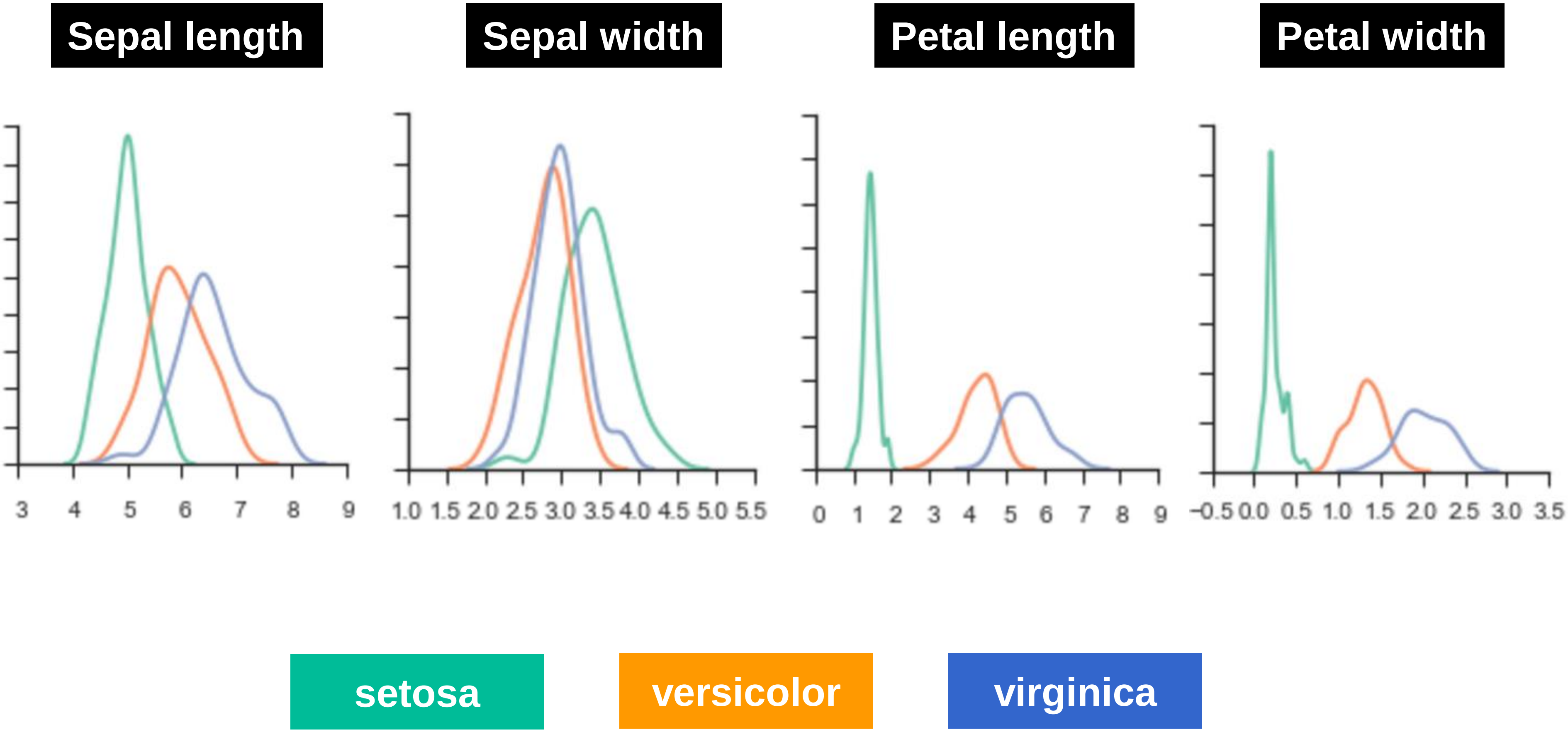
- total of 150 instances
- balanced distribution of the classes

Iris data - overview

Sample number	Sepal length	Sepal width	Petal length	Petal width	Class
1	5.1	3.5	1.4	0.2	**setosa**
2	4.9	3.0	1.4	0.2	**setosa**
…	…	…	…	…	**…**
50	6.4	3.2	4.5	1.5	**veriscolor**
…	…	…	…	…	**…**
150	5.9	3.0	5.1	1.8	**virginica**



Iris data – variable distributions



Iris data summary

- Iris setosa is **linearly** separable from Iris Versicolor and Iris Virginica

- Some overlap between Iris Versicolor and Iris Virginica ➔ perfect classification between the two not possible

- Some redundancy across the four input variables ➔ a good classification model should be achievable with fewer variables

DEMO: Iris data - classification

Resources

- PCA:
  - Python Machine Learning SE, Chapter 5, pages 141 – 154
- PLSR:
  - Video lectures: https://www.youtube.com/playlist?list=PL4C8FE6F00CBBF34A
  - Introduction and examples in R: «The pls Package (Mevik & Wehrens)»