COMP809 Data Mining and Machine Learning

Patricio Maturana-Russel
p.maturana.russel@aut.ac.nz

Department of Mathematical Sciences and Computer Science and Software Engineering Departments, Auckland University of Technology, Auckland, New Zealand

Semester 1, 2024

Contents

- Feature selection
- Feature extraction

Feature selection & Feature extraction

Both techniques aim to reduce the dimensionality of the data.

Feature selection reduces the feature space by removing some of the (non-relevant) features. Example: Chi square test, and ANOVA.

Feature extraction generates new features based on the original dataset. Example: Principal Component Analysis (PCA), and Fourier transformation.

Feature selection & Feature extraction

Dimensionality reduction has many advantages, such as:

- Fewer features potentially imply less complexity.
- Less storage space.
- It speeds up the analyses.
- Model accuracy improves due to less misleading data.
- It removes noise and redundant features.

Feature selection

It selects a subset of features among the set of all features. The idea is to find the optimal set of features.

Also known as *variable selection*, *feature reduction*, *attribute selection*, and *variable subset selection*.

Feature selection approaches

- **Filter methods**: Filtering approaches use a ranking or sorting algorithm to filter out those features that have less usefulness.
- **Wrapper methods**: The attributes subset selection is done using the learning algorithm and generally select features by directly evaluating their impact on the performance of a model.
- **Embedded methods**: they use algorithms that have built in feature selection methods. For instance, Lasso.

Filter methods

Also known as **single factor** analysis, these methods are generally used as a preprocessing step. The method will depend on the type of feature.

	Continuous	Categorical
Continuous	Pearson's correlation	
Categorical	ANOVA	Chi-Square

Pearson's correlation

Pearson's correlation measures **linear relationship** between two variables. Its calculation is given by:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y},$$

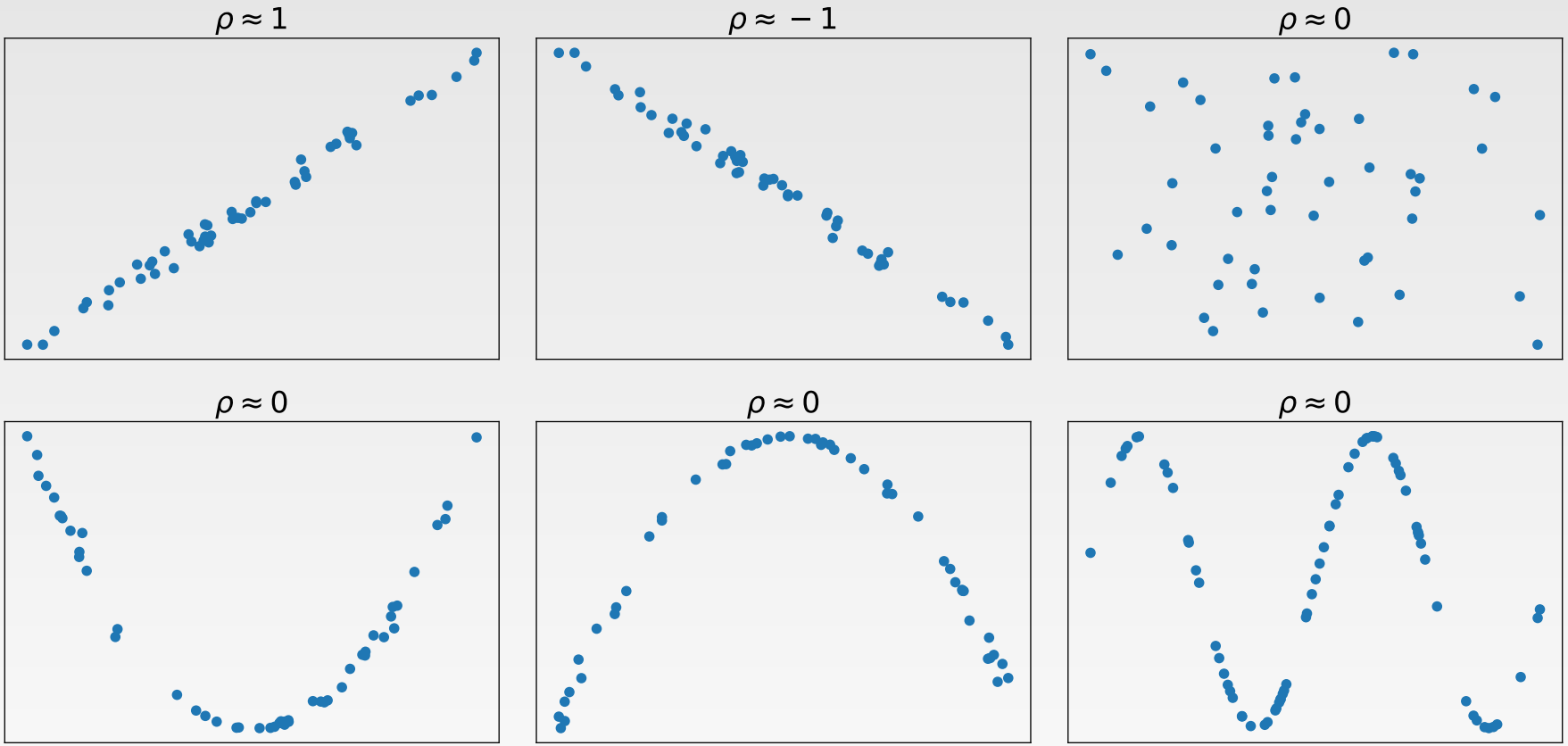
where cov stands for the covariance and σ for the standard deviation.

For NumPy arrays `x` and `y`, it can be calculated as:
`numpy.corrcoef(x,y)`
`scipy.stats.pearsonr(x, y)`

For `Series` objects `x` and `y`, Pandas correlation calculation is given by:
`x.corr(y)`

If two features are linearly correlated, i.e., $\rho \approx -1$ or $\rho \approx 1$, one of them can be discarded.

Pearson's correlation



Note that it can only detect linear relationships.

Pearson's correlation

Interpretation of correlation coefficient:

|Value of Correlation Coefficient|	Relationship Interpretation
0.000 - 0.199	Very weak
0.200 - 0.399	Weak
0.400 - 0.599	Moderate
0.600 - 0.799	Strong
0.800 - 1.000	Very strong

This is the interpretation for the absolute value of the correlation coefficient.

Conclusion: if two variables are correlated, one of them can be removed from further analyses.

ANOVA

The analysis of variance (ANOVA) is used to analyze the differences among means (continuous variable) of groups (categorical variable). The hypotheses to be tested are:

$$\mathrm{H}_0 : \mu_1 = \cdots = \mu_k$$
$$\mathrm{H}_1 : \mu_i \neq \mu_j, \text{for a } i \neq j$$

where μ_i is the mean for the ith group, and k is the number of groups.

Conclusion: if the null hypothesis is not rejected, the categorical variable can be removed from the analysis, since it is not related to the continuous variable.

ANOVA compares two types of variation to test equality of means at each level, partitioning the total variation as:

$$\text{Total Variation} = \text{Between group variation} + \text{Within group variation}$$

If the *between group variation* is significantly greater than *within group variation* then means are not statistically equal.

ANOVA

If H_0 is rejected, you might be interested in knowing which mean is different. For this, you can use the Tukey's test (see www.statology.org).

Assumptions:

1. Independence of the samples.
2. Normally distributed data (less important for large samples due to the Central Limit Theorem).
3. Equality of standard deviations (variability is the same in each group).

ANOVA

A classical data of Michelson on measurements done in 1879 on the speed of light. The data consists of five experiments, each consisting of 20 consecutive 'runs'. The response is the speed of light measurement, suitably coded (km/sec, with 299000 subtracted). Is there a difference between the experiments?

$$\mathsf{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$
$$\mathsf{H}_1 : \mu_i \neq \mu_j, \text{for a } i \neq j \text{ with } i, j = 1, 2, 3, 4, 5.$$

where μ_i is the mean for the ith experiment.

The data is given by

```
exp1 = [850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650, 760, 810, 1000, 1000, 960, 960]
exp2 = [960, 940, 960,  940, 880, 800, 850, 880, 900, 840,  830, 790, 810, 880, 880, 830,  800,  790, 760, 800]
exp3 = [880, 880, 880,  860, 720, 720, 620, 860, 970, 950,  880, 910, 850, 870, 840, 840,  850,  840, 840, 840]
exp4 = [890, 810, 810,  820, 800, 770, 760, 740, 750, 760,  910, 920, 890, 860, 880, 720,  840,  850, 850, 780]
exp5 = [890, 840, 780,  810, 760, 810, 790, 810, 820, 850,  870, 870, 810, 740, 810, 940,  950,  800, 810, 870]
```

ANOVA

One way of performing an ANOVA test in python is as follows:

```
>>> from scipy.stats import f_oneway
>>> f_oneway(exp1, exp2, exp3, exp4, exp5)
    F_onewayResult(statistic=4.2878, pvalue=0.0031)
```

Conclusion: p-value $<$0.05 (significance level), therefore we reject H_0. There is enough statistical evidence to claim that the experiments have different means. Therefore, it is not appropriate to exclude the variable *experiment* from the analysis.

Chi-Square

A chi-square test (also χ^2 test) is a statistical hypothesis test used to examine whether two categorical variables are independent when the sample size is large.

In general terms, the hypotheses are:
H_0: the features are independent.
H_1: the features are not independent.

Conclusion: if the null hypothesis is rejected, one of the features can be excluded from the analysis.

Chi-Square

Suppose we want to find out whether or not vaccination has an effect on a particular form of pneumonia (health outcome is the target variable), i.e., whether or not the variables are independent.

Health Outcome	Not vaccinated	Vaccinated	Total
Sick with pneumococcal pneumonia	23	5	28
Sick with non-pneumococcal pneumonia	8	10	18
Stayed healthy	61	77	138
Total	92	92	184

More information about this test and example on www.ncbi.nlm.nih.gov

If the dataset is a data frame `df`, we can use `pandas.crosstab(df)` to generate the contingency table. Make sure that the variables are categorical. To check the class use `df.dtypes`. For the conversion you can use `df[["var1","var2"]].astype("category")`.

Chi-Square

Roughly speaking, the idea is to compare what is expected, if H_0 is true, with the observed values. In our example, we have that

Health Outcome	Not vaccinated	Vaccinated
Sick with pneumococcal pneumonia	14 (5.79)	14 (5.79)
Sick with non pneumococcal pneumonia	9 (0.11)	9 (0.11)
Stayed healthy	69 (0.93)	69 (0.93)

that contains the cell expected values (E) under H_0 and in parentheses the cell Chi-square values (χ^2), which are calculated as follows:

$$E = \frac{M_R \times M_C}{n} \quad \text{and} \quad \chi^2 = \frac{(O-E)^2}{E},$$

where M_R and M_C represent the row and column marginals for those cells, n the sample size, and O the observed value.

Chi-Square

The χ^2 statistic for the table is calculated by adding the cell χ^2 values. This statistic is compared to the χ^2 distribution with $(n_R - 1) \times (n_C - 1)$ degrees of freedom, where n_R and n_C are the number of rows and columns, respectively.

In this case the χ^2 statistics is 13.65. So, the p-value is given by

```
>>> 1 - scipy.stats.chi2.cdf(13.65, 2)
    0.001086
```

We can perform this analysis directly as follows

```
>>> from scipy.stats import chi2_contingency
>>> data = [[23,8,61], [5,10,77]]
>>> stat, p, dof, expected = chi2_contingency(data)}
```

Conclusion: p-value < 0.05 (significance level), therefore we reject H_0. We reject that the variables are independent, i.e., that we reject the hypothesis that vaccination has no effect on health outcome. Therefore, we cannot exclude the vaccination variable to explain the health outcome.

p-values

Conclusions in terms of the p-value:

$p < 0.001$	Very strong evidence against H_0, supports H_1
$0.001 \leqslant p < 0.01$	Solid evidence against H_0, on supports of H_1
$0.01 \leqslant p < 0.05$	Moderate to good evidence against H_0, on supports of H_1
$0.05 \leqslant p < 0.15$	Little evidence against H_0
$p > 0.15$	None evidence against H_0, but not neccesarily means that H_0 is true

The p-value is the probability, if the null hypothesis is true, of obtaining the observation or an observation more extreme.

Remark: **we never accept H_0!**

Wrapper methods

Wrapper methods measure the usefulness of a subset of feature by training a model on it. The predictive power of a variable is evaluated jointly.

Wrappers can be computationally expensive.

Methods:

- Forward selection.
- Backward selection (or backward elimination).
- Combination of forward and backward selection.

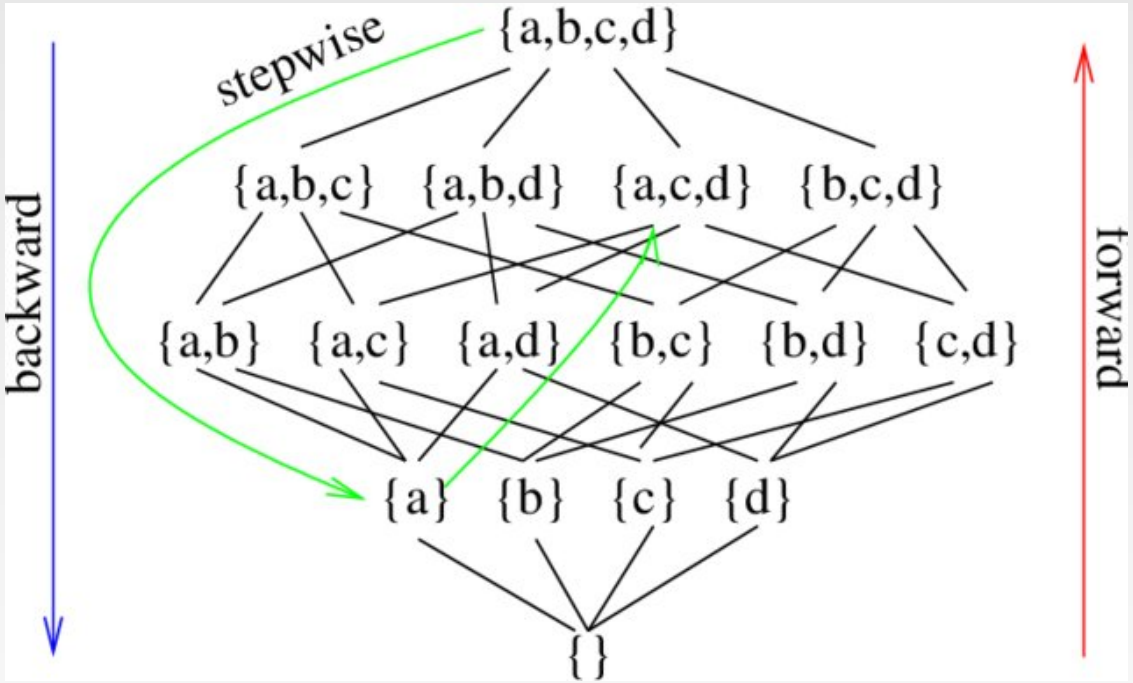
Forward selection

1. Start with an **empty set A** of attributes.
2. At each step an attribute is added and a performance measure is evaluated (for example Correlation map or Information Gain).
3. The attribute that produces the best performance is added to set A.
4. Add each of the remaining attributes to set A and note the attribute with the highest Information Gain (Correlation).
5. This attribute is now added to set A.
6. The entire process is repeated until no more attributes can be added to set A, i.e., at a particular round (iteration) all attributes when added decrease, rather than increase the Information Gain (Correlation).
7. The set A at the end of the process contains the set of non redundant attributes.

Backward selection

1. The set **A** initially consists of the **full set of attributes**.
2. Eliminate (rather than add) at each step the attribute that leads to the highest Information Gain.
3. The process is repeated at the iteration when every attribute that remains in set A leads to a loss of Information Gain.
4. The attributes that remain in set A contain the list of non redundant attributes.

Stepwise selection



Credit: Learning interpretable models by Stefan Rüping

Embedded methods

In an embedded method, feature selection is integrated or built into the classifier algorithm. Less important variables are given lower weight (close to zero).

Example: Consider the linear model

$$y_i = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \cdots + \beta_p \times x_{ip} + \epsilon_i$$

where y is the response, x a predictor, and ϵ the error.

In order to estimate the β parameters, the Lasso regression (Least Absolute Shrinkage and Selection Operator) minimizes

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a penalization parameter.

Some of the β_s are shrunk to exactly zero, excluding those predictors from the model.

Feature extraction

Feature extraction generates new features based on the original dataset.

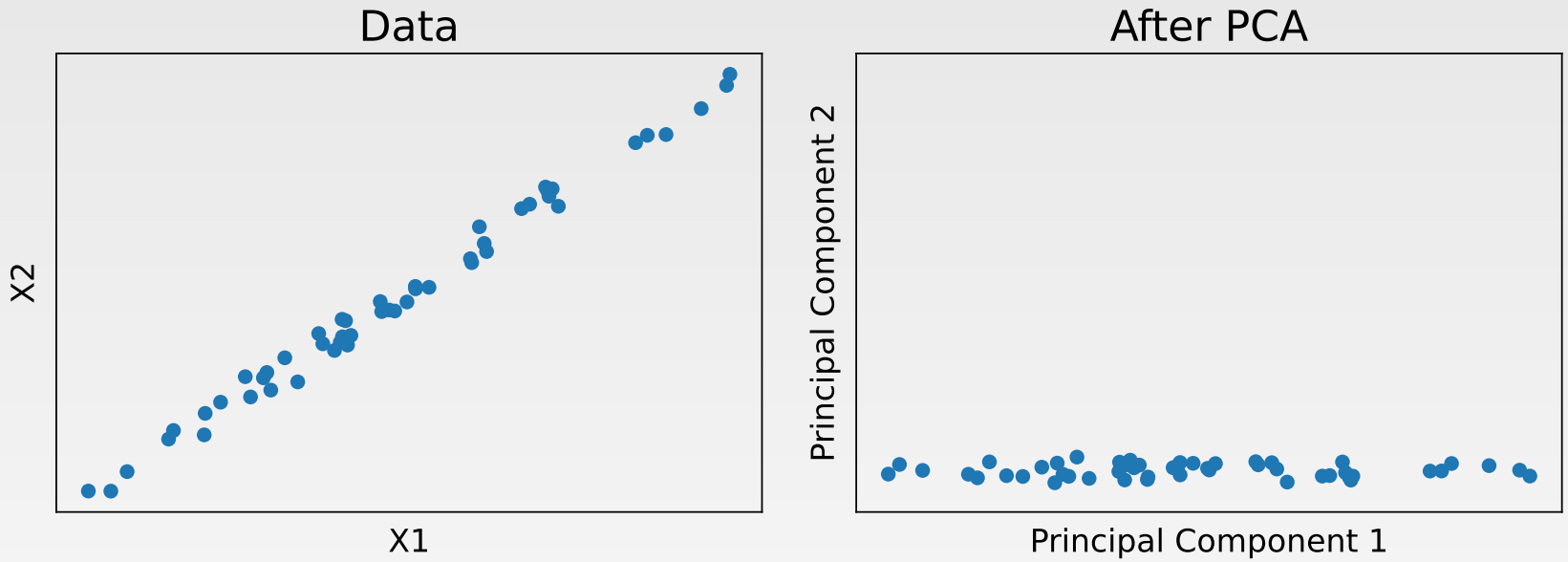
- Principal Component Analysis (studied in this paper).
- Fourier Transformation.
- Vector Quantization.

Principal Component Analysis

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

- It is an exploratory approach to reduce the dataset's dimensionality.
- It finds a linear subspace (passing through the data mean) that results in the smallest (mean square) error between the feature vectors and their projections in the data space.
- It de-correlates feature data via rotation.

Principal Component Analysis



Note that after PCA most of the variability of the data is in the first principal component.

Advantages:

- Removes correlated features.
- Can improves performance and reduce overfitting.
- Reduces dimensionality leads to improve visualization.

Disadvantages:

- Independent variables become less interpretable.
- Data standardization must be before PCA.
  - PCA biased towards features with high variance, leading to false results.

PCA Example

Credit: www.medium.com

The Iris dataset can be downloaded clicking here.

The dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant. More information can be found on www.archive.ics.uci.edu

It can be imported directly from the URL

```
>>> import pandas as pd
>>> url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
          iris.data"
>>> df  = pd.read_csv(url, names=["sepal length","sepal width","petal length",
                                  "petal width","target"])
```

Or imported from the directory as

```
>>> colnames = names=["sepal length","sepal width",
                      "petal length","petal width","target"]
>>> df       =  pd.read_csv("iris.data", names=colnames, header=None)
```

PCA Example

```
>>> print(df)
     sepal length  sepal width  petal length  petal width          target
0             5.1          3.5           1.4          0.2     Iris-setosa
1             4.9          3.0           1.4          0.2     Iris-setosa
2             4.7          3.2           1.3          0.2     Iris-setosa
3             4.6          3.1           1.5          0.2     Iris-setosa
4             5.0          3.6           1.4          0.2     Iris-setosa
..            ...          ...           ...          ...             ...
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica
[150 rows x 5 columns]
```

The exploratory analysis will be omitted.

PCA Example

```
# Separating out the continuous features
>>> features=["sepal length","sepal width","petal length","petal width"]
>>> x = data.loc[:, features].values

# Standardizing the features
>>> from sklearn.preprocessing import StandardScaler
>>> x = StandardScaler().fit_transform(x)

# PCA
>>> from sklearn.decomposition import PCA
>>> pca = PCA(n_components=4)
>>> principalComponents = pca.fit_transform(x)
    # % variability explained by the component
>>> print(pca.explained_variance_ratio_)
    [0.72770452, 0.23030523, 0.03683832, 0.00515193]
```

Percentage explained by the components:

PC1	PC2	PC3	PC4
72.8	23.0	3.7	0.5

Note that the first two components explains 95.8% of the variability.

PCA Example

The new data set to be considered in further analyses is given by

```
>>> PCs = pd.DataFrame(data = principalComponents,
                       columns = ["PC1", "PC2", "PC3", "PC4"])
>>> print(PCs[["PC1","PC2"]])
          PC1       PC2
    0   -2.264542  0.505704
    1   -2.086426 -0.655405
    2   -2.367950 -0.318477
    3   -2.304197 -0.575368
    4   -2.388777  0.674767
    ..        ...       ...
    145  1.870522  0.382822
    146  1.558492 -0.905314
    147  1.520845  0.266795
    148  1.376391  1.016362
    149  0.959299 -0.022284

    [150 rows x 2 columns]
```

PCA Example

```
>>> print(pca.components_) # Linear combination
    [[ 0.52237162 -0.26335492  0.58125401  0.56561105]
     [ 0.37231836  0.92555649  0.02109478  0.06541577]
     [-0.72101681  0.24203288  0.14089226  0.6338014 ]
     [-0.26199559  0.12413481  0.80115427 -0.52354627]]
```

The first two components are calculated as follows

$$\begin{aligned} \text{PC1} &= 0.52 \times sepal\ length - 0.26 \times sepal\ width + 0.58 \times petal\ length + \\ &\quad 0.56 \times petal\ width \\ \text{PC2} &= 0.37 \times sepal\ length + 0.92 \times sepal\ width + 0.02 \times petal\ length + \\ &\quad 0.07 \times petal\ width \end{aligned}$$

PCA2 is mainly composed by *sepal width*, and in an lower degree *sepal length*. The standardized features have been used for these calculations.

PCA Example

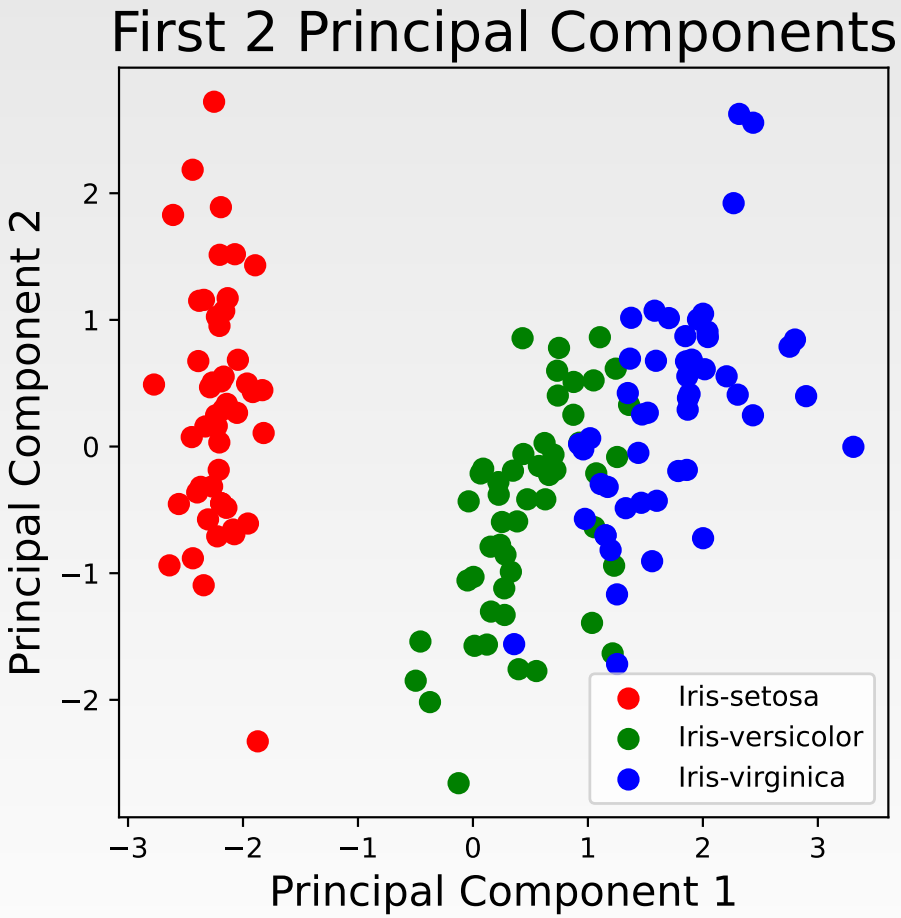
Note that some of the variables are highly correlated. In other words, they contain redundant information. This can cause problems, for instance, in linear regression models.

```
>>> corr_matrix = df[features].corr()
>>> print(corr_matrix)
                sepal length  sepal width  petal length  petal width
sepal length      1.000000    -0.109369      0.871754     0.817954
sepal width      -0.109369     1.000000     -0.420516    -0.356544
petal length      0.871754    -0.420516      1.000000     0.962757
petal width       0.817954    -0.356544      0.962757     1.000000
```

PC1 and PC2 are independent.

PCA Example

- Visualization: 4 dimensional data has been plotted.
- Reduction: these 2 dimensional data can be used for further analysis.
- Classification: patterns can be identified and used for prediction.



Fourier transformation

It is a mathematical function that decomposes functions into frequency components.

In practice, we can apply the Discrete Fourier Transform (DFT) to a time series, obtaining the series in the frequency domain.

In general, working in the frequency domain requires less computational effort.

End