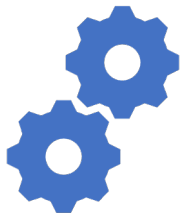
Feature engineering

Alfonso Monaco – INFN Bari

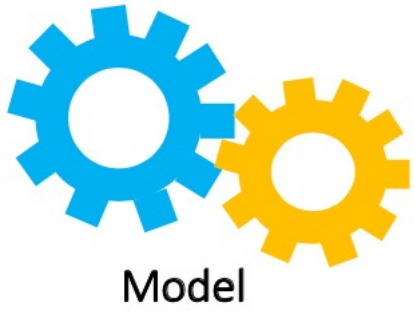


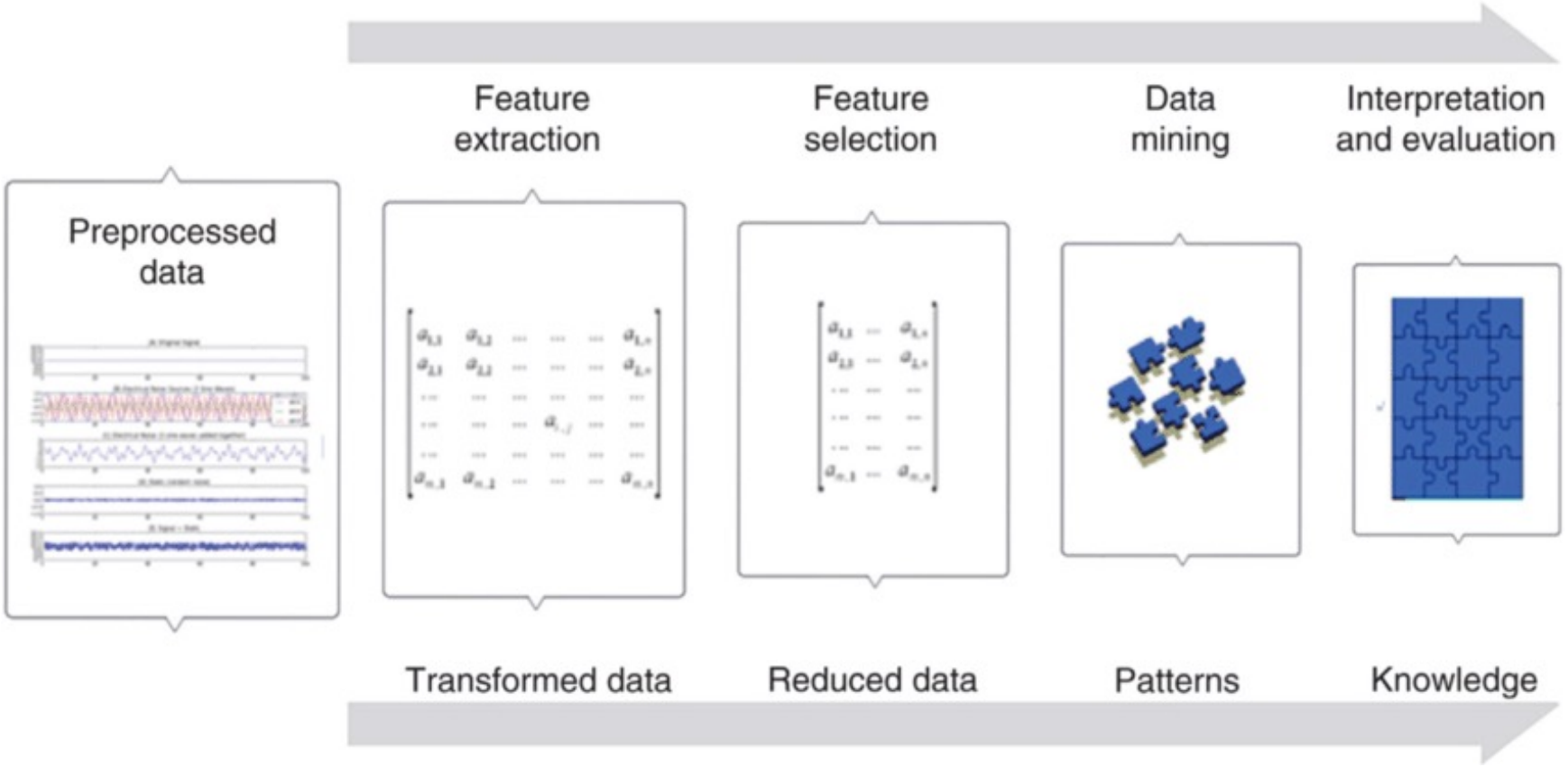
Models and features

Machine learning involves 2 mathematical entities:

Model: a mathematical model of data describes the relationship between different aspects of data. There are mathematical formulas to relate numeric quantities to each other.

Features: a feature is a representation of raw data.




Feature extraction
Feature selection
Data mining
Interpretation and evaluation
Preprocessed data
Transformed data
Reduced data
Patterns
Knowledge

Types of features

Features are tied to the model: some models are more appropriate for some types of features and vice versa;

Feature engineering is the process of formulating the most appropriate features given the data, the model, and the task;

Features can be numerical and categorical.

Numeric data

- Numeric data is already in a format that's easy manageable by mathematical models;

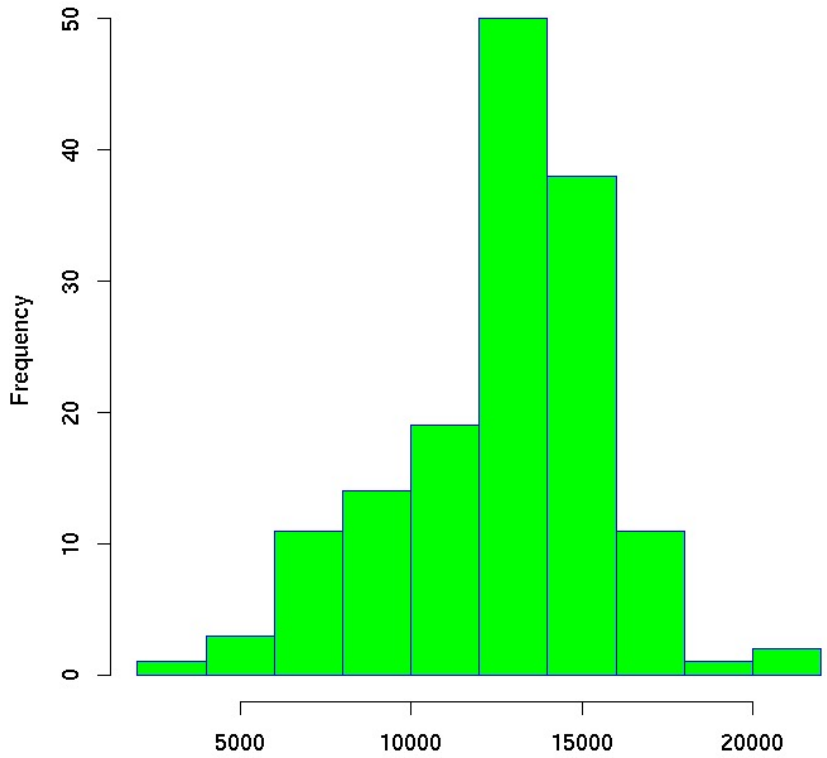
- But this doesn't mean that feature engineering is no longer necessary!

Fancy tricks with numbers

- The first sanity check for numeric data is whether **the magnitude or the scale** matters for the implemented model;
- Next consider **if scaling the features**;
- It's also important to consider the **distribution of numeric features**. For instance, the training process of a linear regression model assumes that features have a Gaussian distribution;
- Multiple features can be **composed together into more complex features**. The hope is that the complex features can more capture important information in raw data;
- In order to reduce the computational expense, It's necessary to prune the input features using automatic **feature selection techniques**.

Quantization or Binning

- Raw counts that span several orders of magnitude are problematic for many models (for example linear model or clustering algorithms as k-means etc…);
- One solution is to contain the scale by quantizing the count;
- We group the feature into bins (like histograms);
- Then we obtained an ordered sequence of bins that represents a measure of intensity.



Quantization or Binning

- We have to decide how wide each bin should be. We can choose through fixed-width or adaptive solution;
- **Fixed-width binning**: each bin contains a specific numeric range that can be custom designed or automatically segmented;
- **Adaptive binning:** with a fixed-width solution many bins could be empty. This problem can be solved by adaptively positioning the bins based on the distribution of the data.
- An example of adaptive binning is **quantiles** that divide the data into equal portions.

Feature scaling or normalization

Such models as linear regression logistic regression, neural networks etc… are affected by the scale of the input;

If the model is sensitive to the scale of the input features, feature scaling could help. We can consider three types of scaling or normalization:

- Min-Max scaling;
- Variance scaling;
- L^2 normalization.

Min-Max scaling

- Let's a feature x

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Variance scaling or standardization

- Let's a feature x

$$\tilde{x} = \frac{x - \mathrm{mean}(x)}{\sqrt{var(x)}}$$

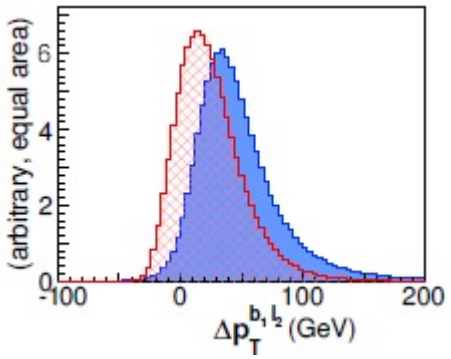
L^2 normalization

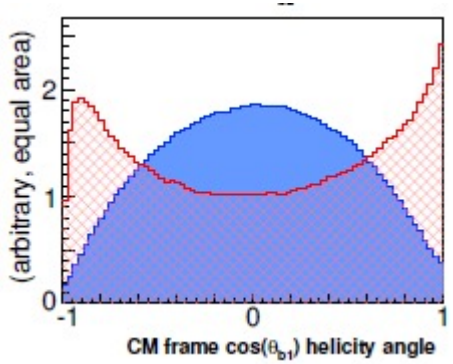
- Let's a feature x

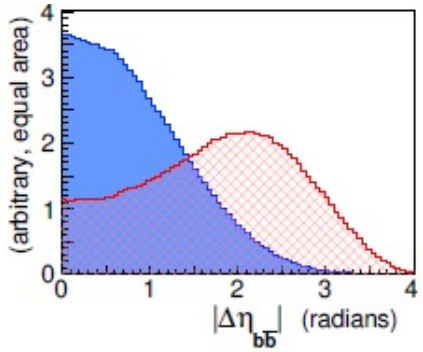
$$\tilde{x} = \frac{x}{\sqrt{x_1^2 + x_2^2 + \ldots + x_m^2}}$$

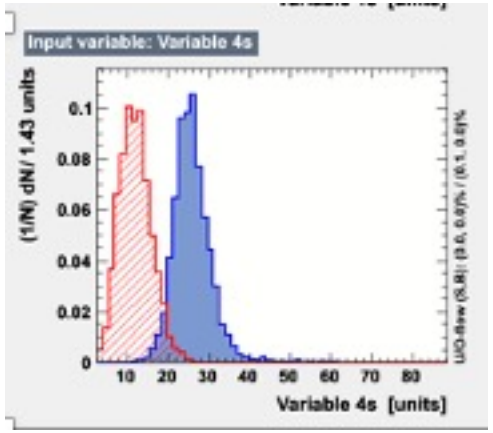
Important: feature scaling always divides the feature by a constant. Therefore, it does not change the shape of the single feature distribution.

Feature distributions









Feature selection

Feature selection is a procedure that selects a subset of relevant and informative features to be used in a model;

The central premise of feature selection is that the data contain some feature that are either redundant or irrelevant, and that can thus be removed without incurring much loss of information;

A large number of features can lead several disadvantages such as computational burden or the model overfitting.

Feature selection

In general feature selection methods include three different techniques can also be used in series:

Filter methods;

Wrapper methods;

Embedded methods.

Filter methods

- Filtering techniques preprocess features to remove ones that are unlikely to be useful for the model;
- For example, one could compute the correlation between each feature and filter out the feature that fall below a threshold

$$r(a,b) - \frac{\sum_{i=1}^{N}(a_i - \bar{a})\sum_{i=1}^{N}(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2 \sum_{i=1}^{N}(b_i - \bar{b})^2}}$$

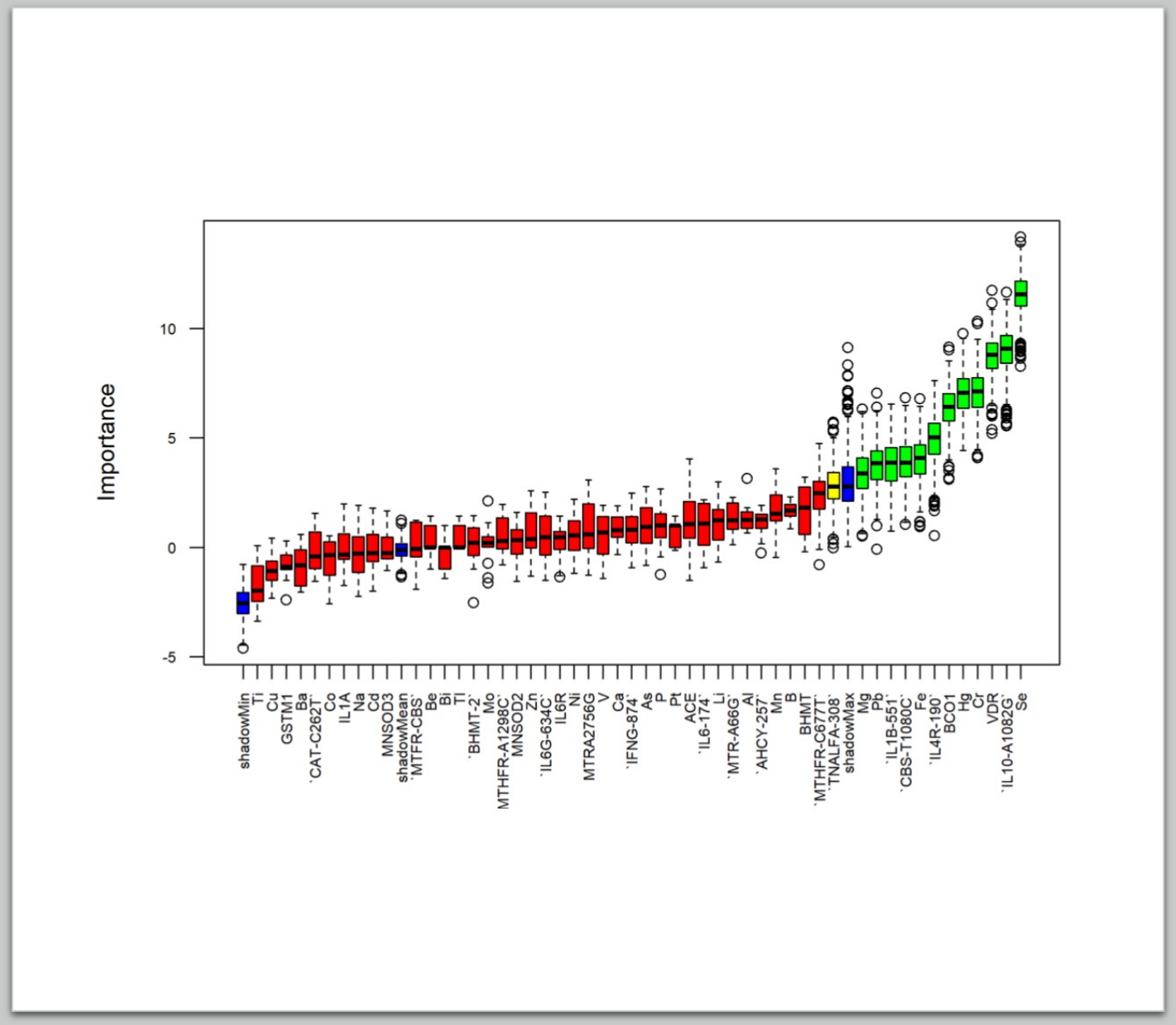
The coefficient r has a value between 1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

Wrapper methods

- Wrapper methods use a predictive model to score each feature subset used to train a model, which is tested on a control set;
- In general these methods are computationally intensive;
- An example of wrapper method is stepwise regression/classification in which the choice of predictive variables is carried out by an automatic procedure;
- In this procedure, at each step, a variable is considered for addition to or subtraction from the set of explanatory features by means of some specified evaluation criterion;

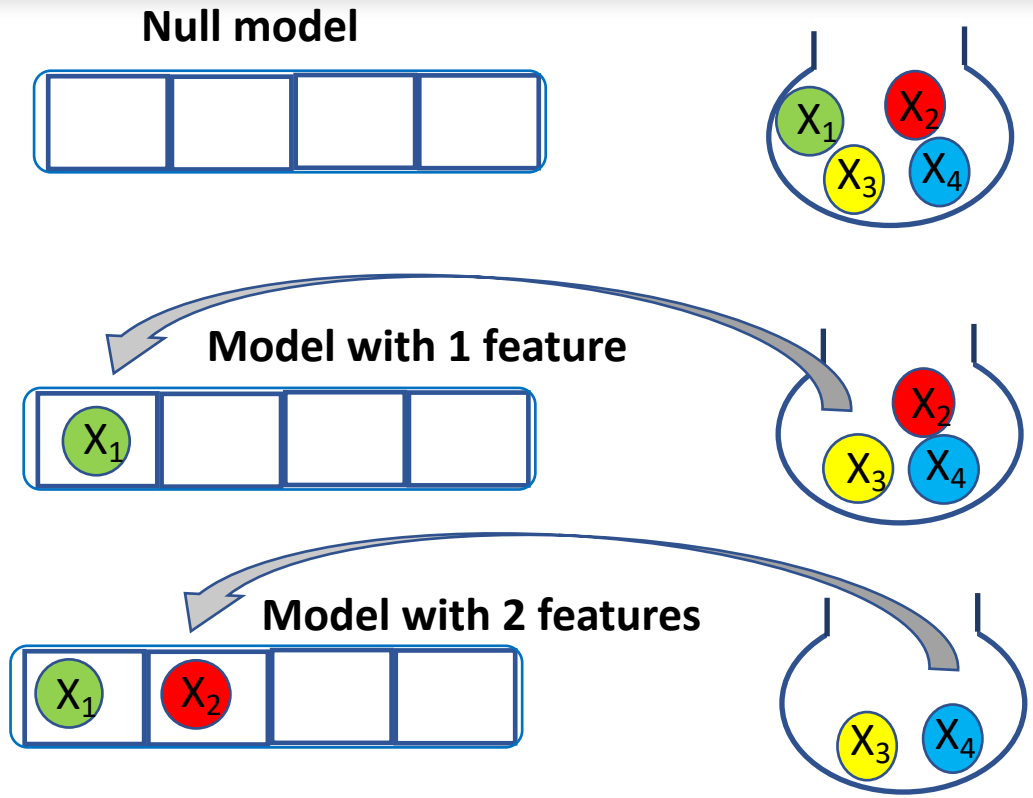
Example of wrapper method

- Boruta uses Random Forest as a model



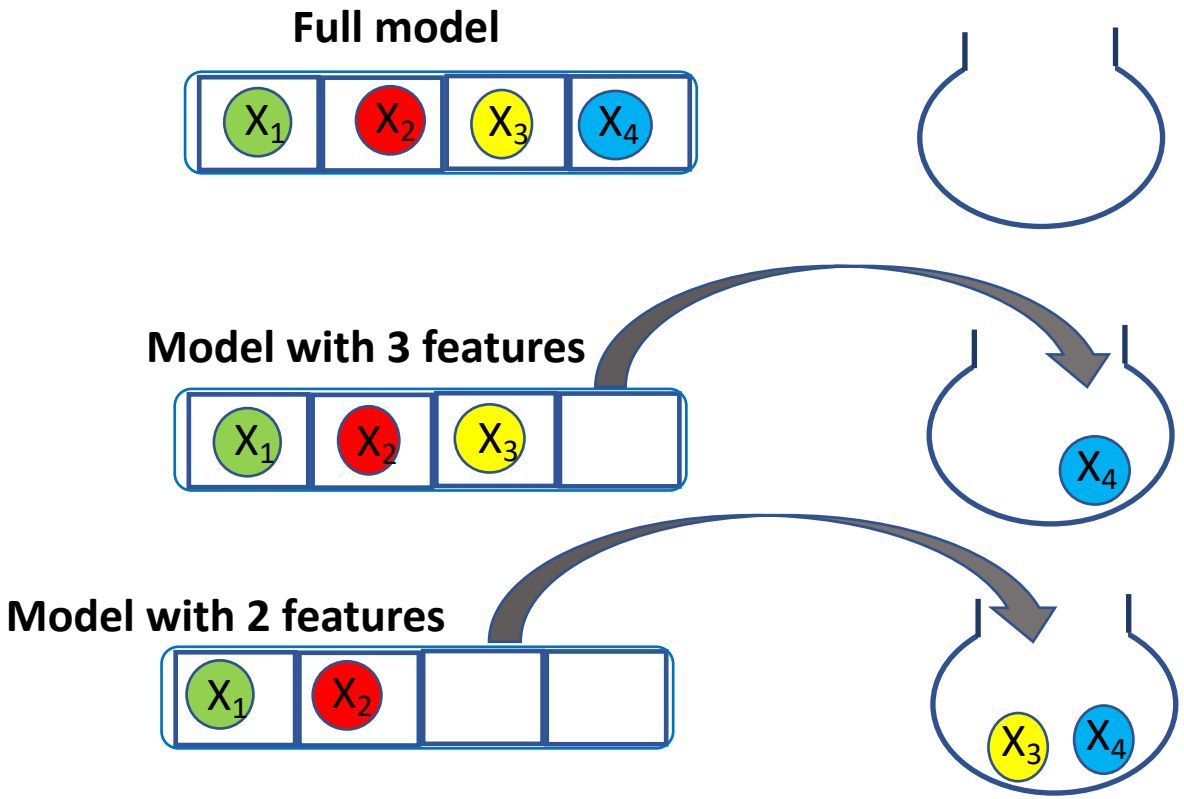
Types of stepwise regression

- **Forward Selection:** starts with no variables in the model, evaluates the addition of each feature using a chosen criterion, adds the feature whose inclusion gives the most statistically significant improvement of the fit, and repeats this process until there is no statistically significant improvement;



Types of stepwise regression

- **Backward Elimination:** starts with all candidate features, evaluates the deletion of each variable using a chosen criterion, deletes the variable whose loss gives the less statistically significant deterioration of the model fit, and repeats this process until no further variables can be deleted at a prespecified significance threshold.



Embedded methods

- Embedded methods try to combine advantages of both Filter and Wrapper methods, but unlike these they do not separate the learning process from the feature selection process.

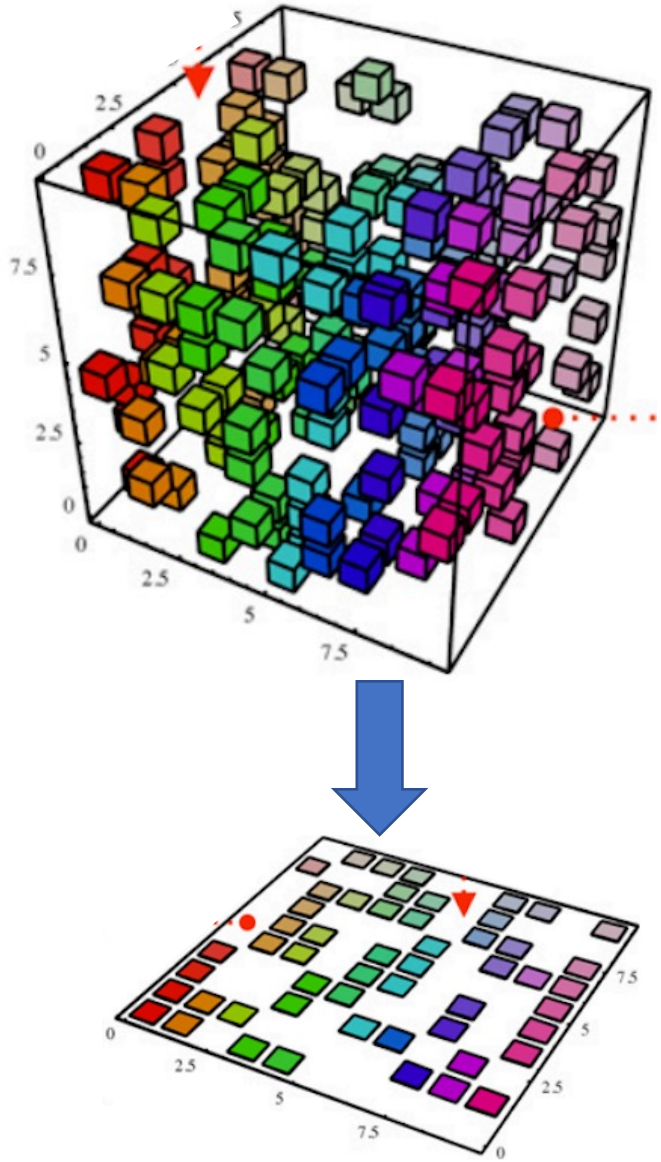
- LASSO, for instance, is an embedded feature selection method (embedded in the regression problem).

Dimension reduction

Why is dimension reduction something that we are interested in doing?

One reason may be for the sake of compressing data. If a dataset is particularly large, it may be useful to reduce the dimensions of the data;

A more interesting reason for dimension reduction is that it provides insights into the underlying structure of our data and the ways that different attributes relate to each other.



Example 1: Pac-Man

- Pac-Man is a little circular creature on a screen who likes to eat little dots and fruits;
- He lives in a maze that he has to navigate with only two sets of directions to move in: up/down and left/right;
- There are some monsters who try to chase Pac-Man and kill him.

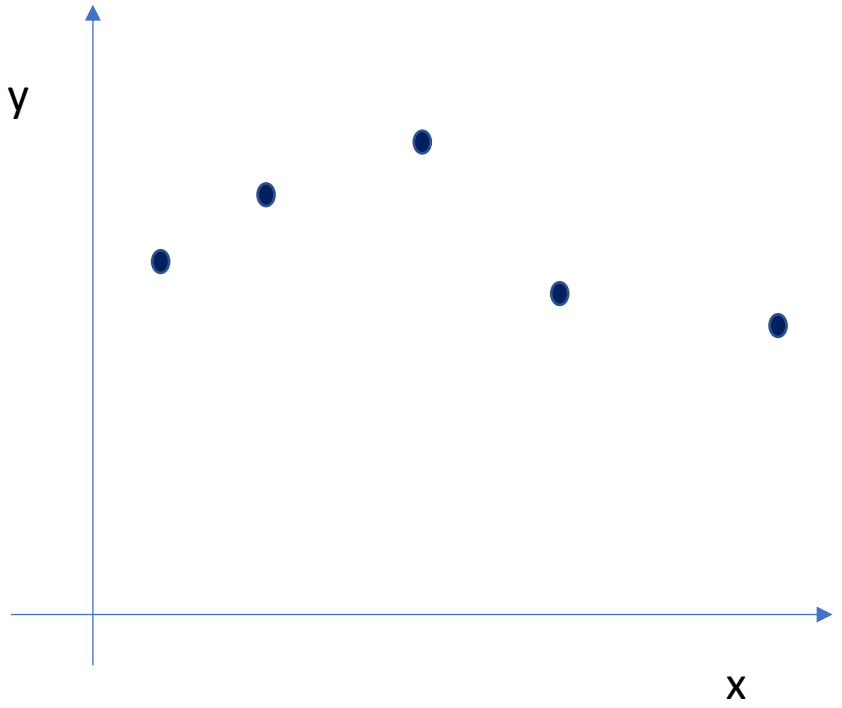
Example 1: Pac-Man

- Pac-Man's position can be fully described by two numbers:

1. how far he is from the left side of the screen;
2. how far he is from the top of the screen.

- If we know those two numeric measurements, then there is only one unique place on the screen where he could be;

- So, if we wanted to collect data on where Pac-Man was over time, we would be able to collect a two-dimensional dataset;



Example 1: Pac-Man

- Now suppose that the dimensions of this dataset are 3:

1. how far Pac-Man is from the left side of the screen;
2. how far Pac-Man is from the top of the screen;
3. how far Pac-Man is from the blue monster that is chasing him.

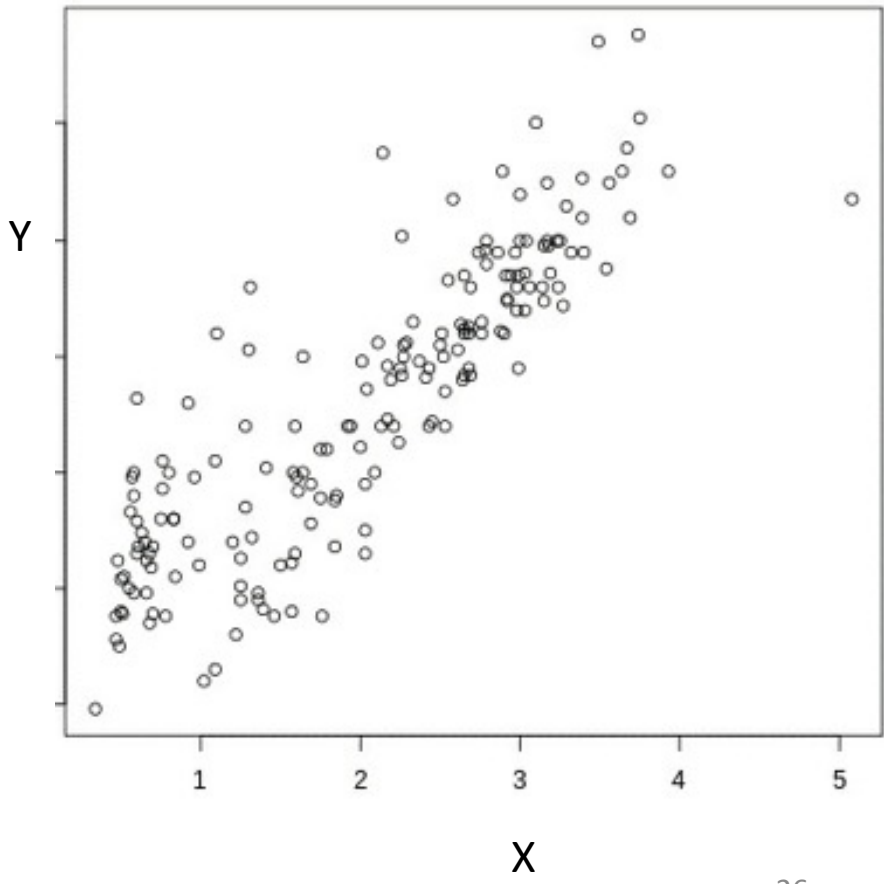
- This is a three-dimensional dataset; however, we can have complete knowledge of Pac-Man's location with only the information contained in the first two dimensions;

- Then the third dimension is unnecessary.

Example 2

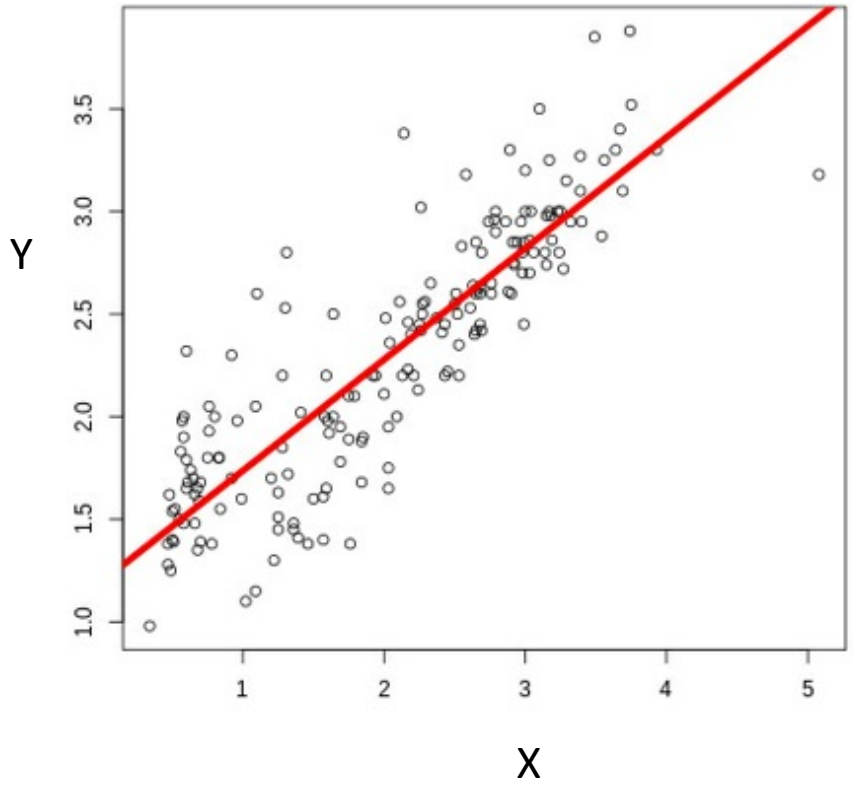
- Consider the following scatterplot of the two-dimensional data;

- After plotting the data, we observe that there appears to be a strong correlation between X and Y.



Example 2

- We can draw a line on the plot that represents this correlation;
- the red line follows the geometric shape of our data quite closely;
- If we wanted a concise way to describe the points, we could simply say what point on the red line they are closest to.
- This would not be a perfect description of the data. However, describing this data using only the red line is a reasonable approximation to the actual data.

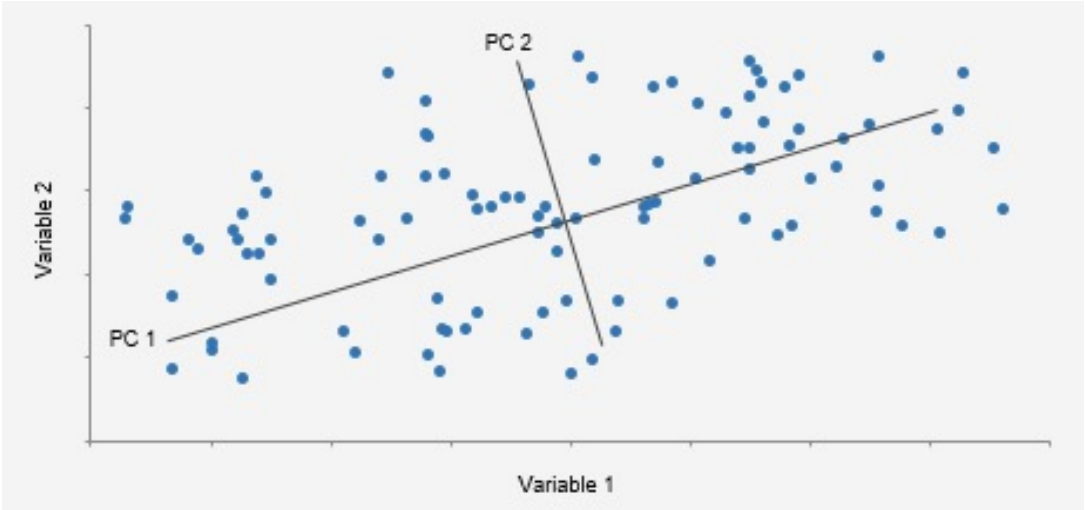


Example 2

- If we describe each observation using the point on the red line that it is closest to, then what we have accomplished is dimension reduction;
- We started with a dataset that requires two measurements to describe each observation and found a way to describe each observation using only one point;
- This is the basic idea of every dimension reduction strategy.

Principal Component Analysis (PCA)

- PCA is a dimension reduction method;
- This is a very common technique used by researchers in a wide variety of fields.



Initial notions: Variance

- **Variance:** In general, the variance of a variable gives us an idea of how widely that variable is spread out;
- It is a measure of the deviation from the mean for points in one dimension.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

Initial notions: Covariance

- **Covariance**: Covariance is variance that is measured for two different variables together. It measures the extent to which their dispersion matches. In other words, it measures the extent to which if one is high, the other is also high, and how high each of them is expected to be;

- It is a measure of how much each of the dimensions vary from the mean with respect to each other;

- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions.

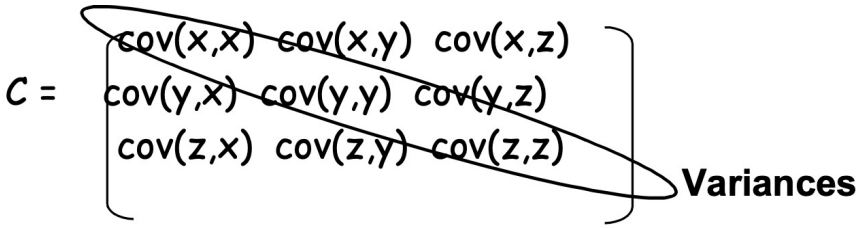
Initial notions: Covariance

- The covariance between one dimension and itself is the variance

$$\text{covariance}(X,Y) = \frac{\sum_{i=1}^{n} (\overline{X_i} - X)\,(\overline{Y_i} - Y)}{(n-1)}$$

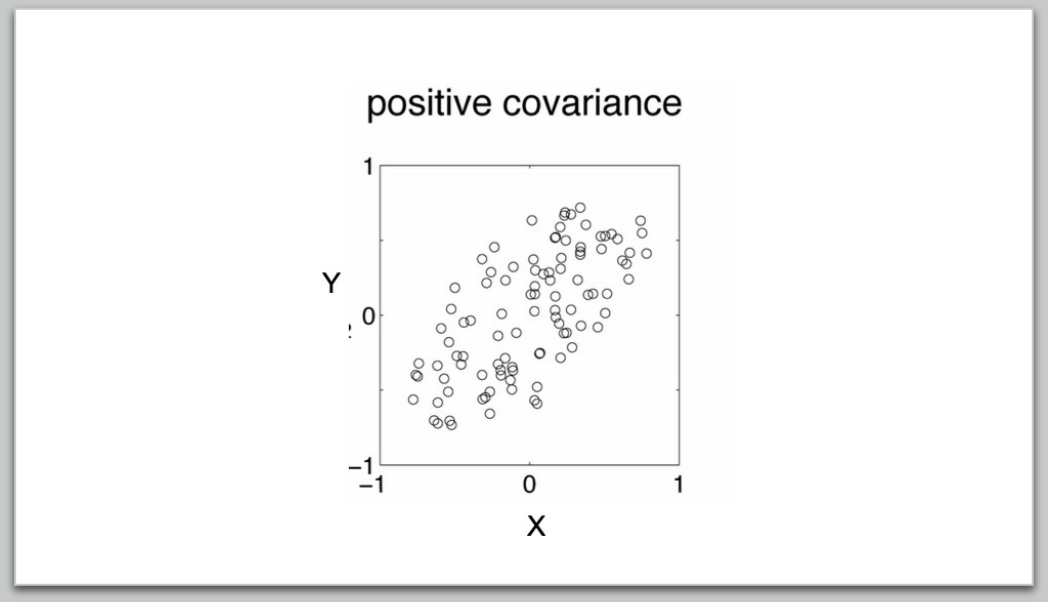
Initial notions: Covariance

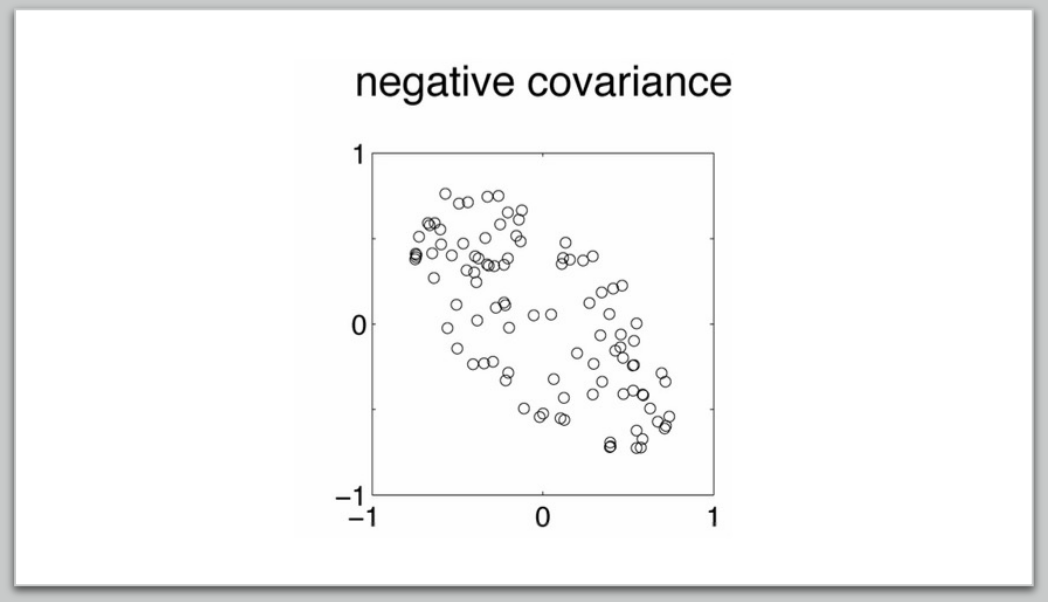
- If you have a 3-dimensional data set (x,y,z), then you could measure the covariance between the x and y dimensions, the y and z dimensions, and the x and z dimensions.



What is the interpretation of covariance calculations?

- A positive value of covariance indicates both dimensions increase or decrease together;
- A negative value indicates while one increases the other decreases, or vice-versa;
- If covariance is zero: the two dimensions are independent of each other.





Covariance

- Why bother with calculating covariance when we could just plot the 2 values to see their relationship?

Covariance calculations are used to find relationships between dimensions in high dimensional data sets (usually greater than 3) where visualization is difficult.

Initial notions: Eigenvectors and Eigenvalues

When we have a square matrix such as a covariance matrix, there are certain special vectors we can calculate called **eigenvectors.**

Each eigenvector has an associated a value called an **eigenvalue**.

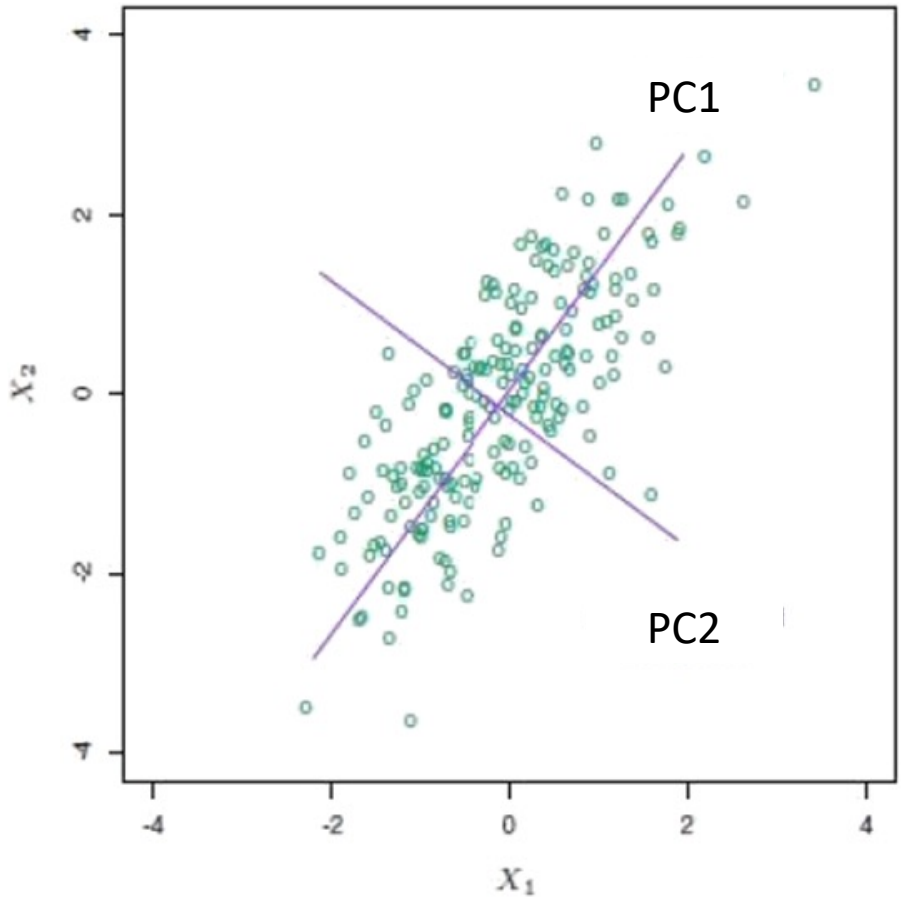
For our purposes, the most important thing to know about eigenvectors is that they express the directions of maximum variance in our data.

The most important thing to know about eigenvalues is that they indicate which eigenvectors are the most important.

PCA

- **principal components analysis (PCA)** is a technique that can be used to simplify a dataset;
- It is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component);
- the second greatest variance on the second axis, and so on;
- PCA can be used for reducing dimensionality by eliminating the later principal components.

PCA



PCA

- To accomplish PCA, we will take the covariance matrix of our data, and then find its eigenvectors;
- The eigenvectors of the covariance matrix are called principal components;
- The principal components enable us to re-express the data in different terms and different numbers of dimensions.

How PCA works

- You suppose we have an initial matrix with 10 features and 8 observations;

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
O1										
O2										
O3										
O4										
O5										
O6										
O7										
O8										

How PCA works

- Then we build the covariance matrix

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
F1										
F2										
F3										
F4										
F5										
F6										
F7										
F8										
F9										
F10										

How PCA works

- The covariance matrix of the initial dataset will have 10 eigenvectors;
- An eigenvalue is associated with each eigenvector. The 10 eigenvectors are ordered by the value of the associated eigenvalue.

	EV1	EV2	EV3	EV4	EV5	EV6	EV7	EV8	EV9	EV10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

evalue1
evalue3
evalue3
evalue4
evalue5
evalue6
evalue7
evalue8
evalue9
evalue10

How PCA works

- So, each element of the eigenvector is a coefficient in an equation to generate a new principal component.
- Each principal component is a linear combination of the original features:

$$PC1 = e_{1,1} * f_1 + e_{1,2} * f_2 + e_{1,3} * f_3 + e_{1,4} * f_4 + e_{1,5} * f_5 + e_{1,6} * f_6 + e_{1,7} * f_7 + e_{1,8} * f_8 + e_{1,9} * f_9 + e_{1,10} * f_{10}$$

.

.

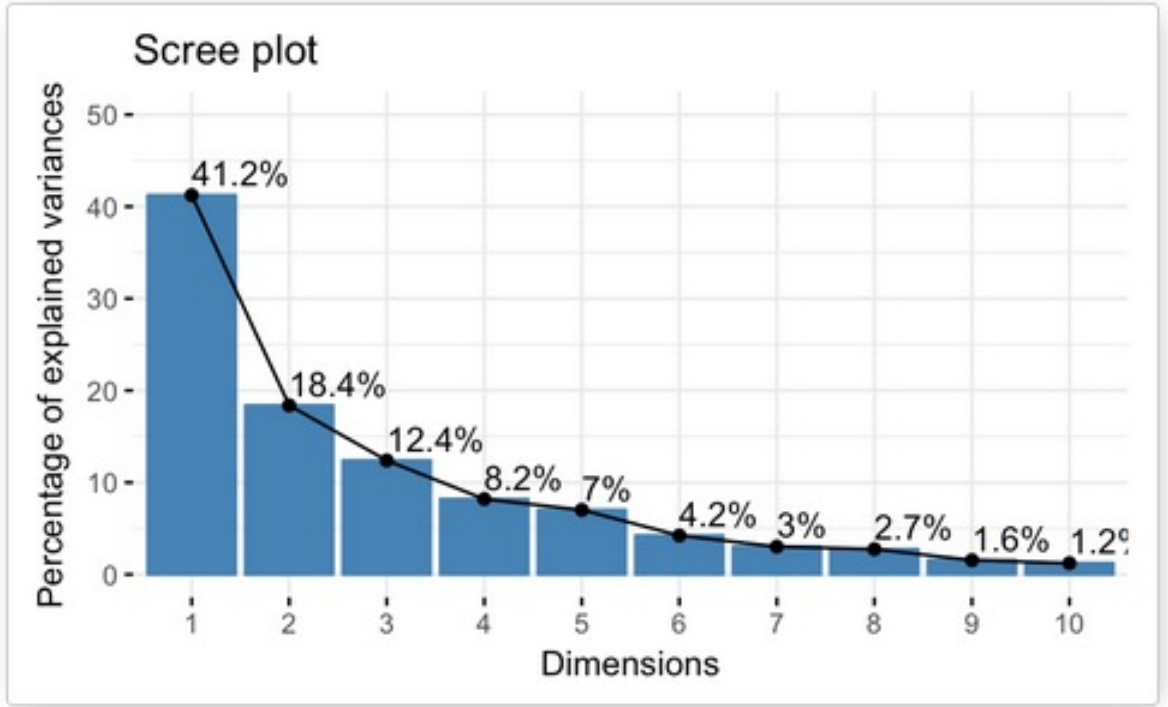
$$PC10 = e_{10,1} * f_1 + e_{10,2} * f_2 + e_{10,3} * f_3 + e_{10,4} * f_4 + e_{10,5} * f_5 + e_{10,6} * f_6 + e_{10,7} * f_7 + e_{10,8} * f_8 + e_{10,9} * f_9 + e_{10,10} * f_{10}$$

How PCA works

- PCA enables us to do dimension reduction. Instead of re-expressing the data in terms of 10 new dimensions defined by the eigenvectors, we can select only the most important of these new dimensions;
- This is simple because the importance of each eigenvector is measured by its corresponding eigenvalue;
- We used the screen plot.

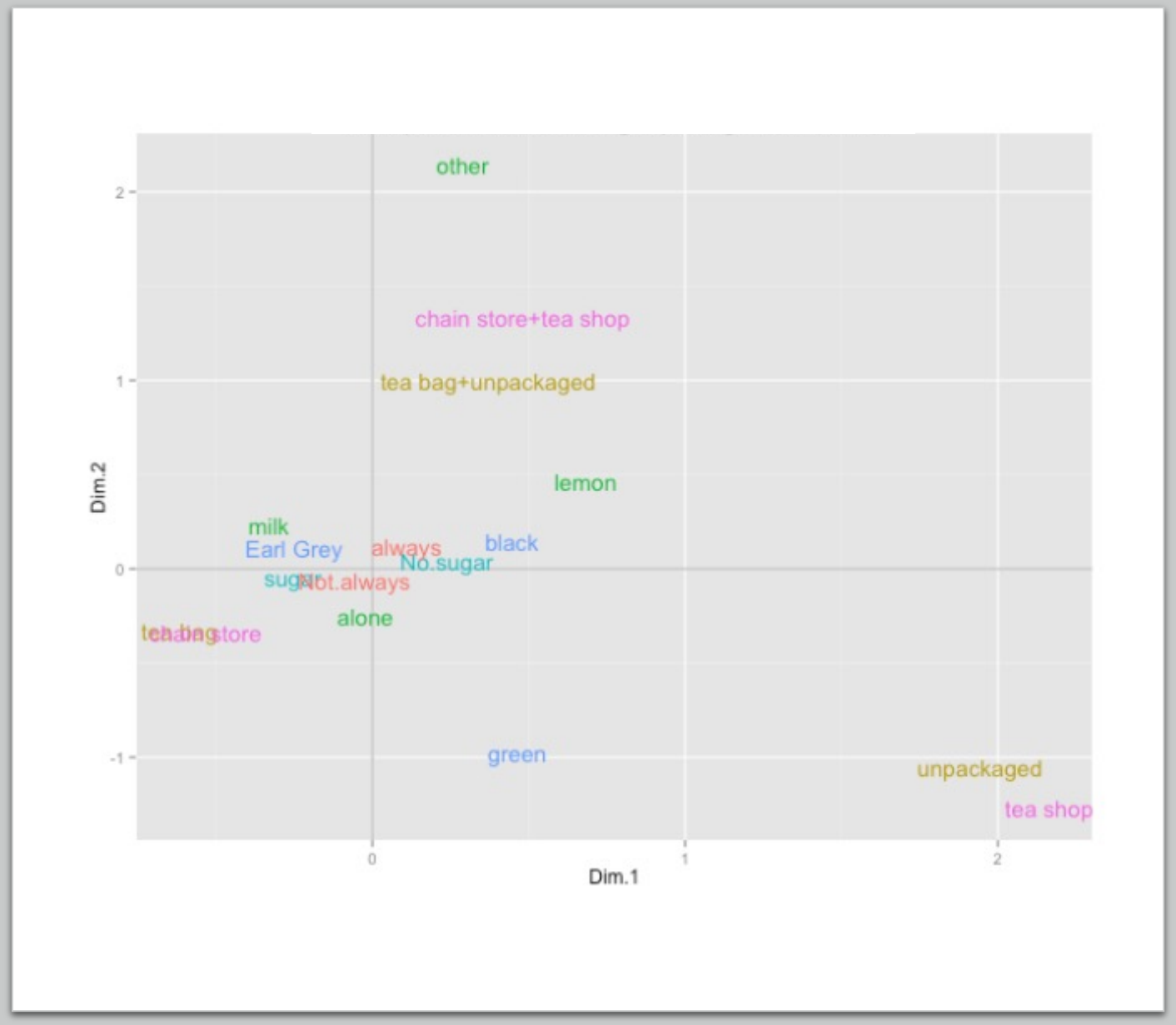
Screen plot

- A scree plot is a simple line segment plot that shows the eigenvalues of a matrix represented in order from highest to lowest, in order to indicate the relative importance of their associated eigenvectors;
- We will use the scree plot to decide which eigenvectors (that is, which dimensions) are the most important.
- IMPORTANT: before using PCA the data must be normalized.



Multiple Corresponde e Analysis (MCA)

- PCA for categorical data is called Multiple Correspondence Analysis (MCA).



Linear Discriminant Analysis (LDA)

- Both LDA and PCA are linear transformation techniques: LDA is a supervised whereas PCA is unsupervised.
- In contrast to PCA, LDA attempts to find a feature subspace that maximizes class separability.