Lecture notes on Regression: Markov Chain Monte Carlo (MCMC)

Dr. Veselina Kalinova, Max Planck Institute for Radioastronomy, Bonn
"Machine Learning course: the elegant way to extract information from data", 13-23 Ferbruary, 2017

1 Overview

Supervised Machine Learning (SML): having input variables (X) and output variables (Y), adopt an algorithm to understand the mapping function: Y=f(X).

Goal of SML: to find a function f(X), which can predict well the new output values (Y') for a given set of input values (X').

- Regression problem: the output value is a *real value*, i.e., "height" or "weight"
- Classification problem: the output value is a *category*, i.e., "red" or "blue"

2 Regression Analysis

2.1 Linear regression

Example: a) line-fitting

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1}$$

where i=1,...,n is the particular observation; β_0 and β_1 - linear parameters, ϵ_i - error.

b) parabola-fitting

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{2}$$

where i=1,...,n is the particular observation; β_0, β_1 and β_2 - linear parameters, ϵ_i - error (still linear regression, because the coefficients β_0, β_1 and β_2 are linear, although there is a quadratic expression of x_i).

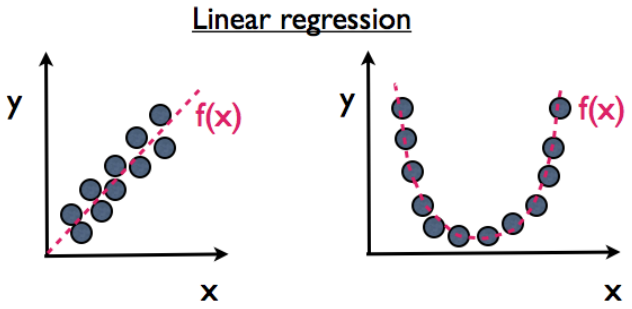

Figure 1: Linear Regression for a given data set x. *left:* line-fitting. *Right:* parabola-fitting.

2.2 Ordinary least squares

We define the residual e_i as the difference between the value of the dependent variables, predicted by the model y_i^{MOD} and the true value of the dependent variables, y_i^{OBS}, i.e.,

$$e_i = y_i^{OBS} - y_i^{MOD} \tag{3}$$

One way to estimate the residuals is through "ordinary least squares" method, which minimaze the *sum of the squared residuals (SSE)*:

$$SSE = \sum_{i=1} ne_i^2 \tag{4}$$

The the mean square error of regression is calculated as

$$\sigma_\epsilon^2 = SSE/dof, \tag{5}$$

where $dof = (n - p)$ with n-number of observations, and p-number of parameters or $dof = (n - p - 1)$ if intersept is used.

2.3 Best fit of function

The best fit of a function is defined by the value of the "chi-square":

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i^{OBS} - y_i^{MOD}]^2}{\epsilon_{y_i}^2}, \tag{6}$$

where our data/observations are presented by y_i^{OBS} with error estimation ϵ_{y_i}, and model function y_i^{MOD}.

It is aslo necessary to know the number of degrees of freedom of our model ν when we derive the χ^2 , where for n data points and p fit parameters, the number of degrees of freedom is $\nu = n - p$. Therefore, we define a reduced chi-square χ_ν^2 as a chi-square per degree of freedom ν:

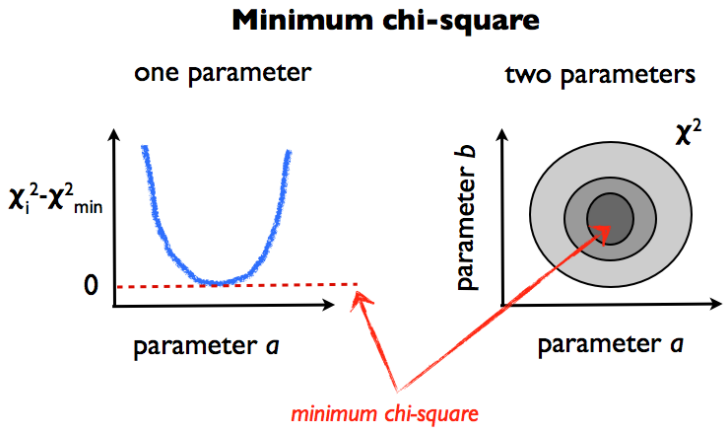

Figure 2: The minimum chi-square method aims to find the best fit of a function. *left:* for one parameter, *right:* for two parameters

$$\chi^2_\nu = \chi^2/\nu, \tag{7}$$

where $\nu = (n - m)$ with n - number of measurements, and p - number of fitted parameters.

- $\chi^2_\nu < 1 \rightarrow$ over-fitting of the data
- $\chi^2_\nu > 1 \rightarrow$ poor model fit
- $\chi^2_\nu \simeq 1 \rightarrow$ good match between data and model in accordance with the data error

2.4 The minimum chi-squared method

Here the optimum model is the satisfactory fit with several degrees of freedom, and corresponds to the minimusation of the function (see Fig.2, left). It is often used in astronomy when we do not have realistic estimations of the data uncertainties.

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i^2}, \tag{8}$$

where O_i - observed value, and E_i - expected value.

If we have the χ^2 for two parameters, the best fit of the model can be represented as a contour plot (see Fig.2, right):

3 Markov Chain Monte Carlo

3.1 Monte Carlo method (MC):

- Definition:
  "MC methods are computational algorithms that rely on repeated random sampling to obtain numerical results, i.e., using randomness to solve problems that might be deterministic in principle".

- History of MC:
  First ideas: G. Buffon (the "needle experiment", 18th century) and E. Fermi (neutron diffusion, 1930 year)
  Modern application: Stanislaw Ulam and John von Neumann (1940), working on nuclear weapons projects at the Los Alamos National Laboratory
  The name "Monte Carlo": chosen as a secret name for the nuclear weapons projects of Ulam and Neumann; it is named after the casino "Monte Carlo" in Monaco, where the Ulam's uncle used to play gamble. MC reflects the randomness of the sampling in obtaining results.

- Steps of MC:
  a) define a domain of possible inputs
  b) generate inputs randomly from a probability function over the domain
  c) perform deterministic computation on the inputs (one input - one output)
  d) aggregate (compile) the results

- Perfoming Monte Carlo simulation (see Fig.3):

  Area of the triangle, $A_t = \frac{1}{2}x^2$
  Area of the square, $A_{box} = x^2$

  Therefore, $\frac{1}{2} = \frac{A_t}{A_{box}} \Rightarrow A_{box} = \frac{1}{2}A_t$.

  We can define the ratio between any figure inside the square box by random sampling of values.

  $\frac{16}{30} \sim \frac{1}{2}$ by counting the randomly seeded points

  $\frac{1}{2} \sim \frac{counts\,in\,triangle}{counts\,in\,box}$

  Monte Carlo algorithm is random - more random points we take, better approximation we will get for the area of the triangle (or for any other area imbeded in the square) !

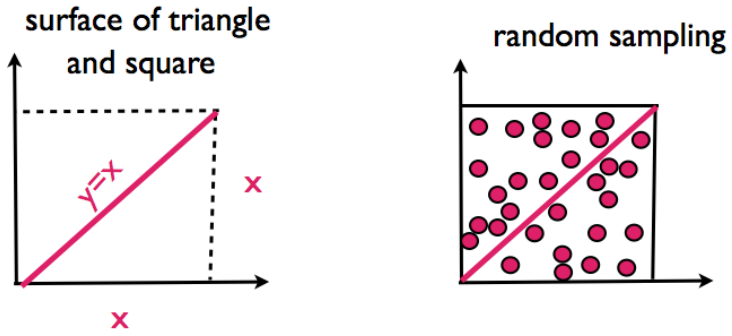

Figure 3: Example for Monte Carlo simulation - random sampling of points (right) to find the surface of a triangle inside a square (left), i.e., the ratio between the counts of the points wihin the two regions will give us the ratio of their surfaces. Our estimation will be more accurate if we increase the numbers of the points for our simulation.

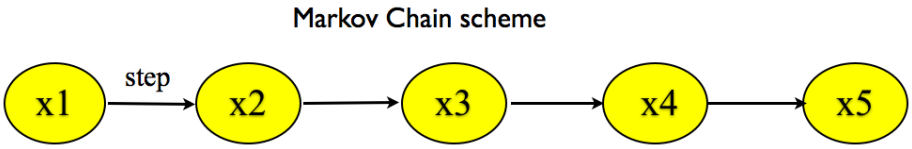

Figure 4: Representation of a Markov Chain.

3.1.1 Markov Chain

- **First idea**: Andrey Markov, 1877

- **Definition**: If a sequence of numbers follows the graphical model in Fig. 4, it is a "Markov Chain". That is, the probability p_i for a given value x_i for each step "i":

$$p(x_5|x_4, x_3, x_2, x_1) = p(x_5|x_4) \tag{9}$$

The probability of a certain state being reached depends only on the previous state of the chain!

- **A descrete example for Markov Chain**:
We construct the Transition matrix **T** of the Markov Chain based on the probabilities between the different state X_i, where i is the number of the chain state:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

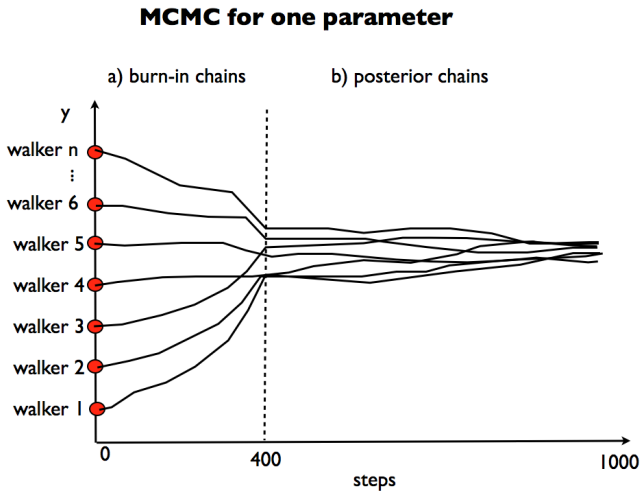

Figure 5: Markov Chain Monte Carlo analysis for one fitting parameter. There are two phases for each walker with an initial state: a) burn-in chain and b) posterior chain.

Let's take a starting point X_0 with initial probabilities $X_0 = (0.5, 0.2, 0.3)$. The next step X_1 will evolve as $X_1 = X_0 \times \mathbf{T} = (0.2, 0.6, 0.2)$ and the system will converge to $X_{converged} = (0.2, 0.4, 0.4)$.

Additional two conditions have to be applied in the evolution of the system, the chains have to be:
a) *Irreducible* - for every state X_i, there is a positive probability of moving to any other state.
b) *Aperiodic* - the chain must not get trapped in cycles.

- **Phases of the Markov Chain (see Fig.5):**
  a) *"burn-in"* chain - throwing some initial steps from our sampling, which are not relevant and not close to the converging phase (e.g., we will remove some stuck walkers or remove "bad" starting point, which may over-sample regions with very low probabilities)

  b) *posterior* chain - the distribution of unknown quantities treated as a random variables conditional on the evidence obtain from an experiment, i.e. this is the chain after the burn-in phase, where the solution settles in an equalibrium distribution (the walkers oscillate around a certain value)

4 Bayes' theorem

- **First idea**: Thomas Bayes, 1701-1761

- **Importance**: Bayes's theorem is fundamental for Theory of Probability as the Pythagorean theorem for the Geometry.

- **Definition**: it can be expressed as the following equation of probability functions

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{10}$$

where A and B are events, $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are probabilities of observing event A and event B without regard to each other; $P(A)$ is also called "prior" probability, while $P(B)$ - "evidence" or "normalisation" probability
- $P(A|B)$ is a conditional probability, i.e., the probability of observing event A given that the event B is true (or "posterior" probability)
- $P(B|A)$ is a conditional probability, i.e., the probability of observing event B given that the event A is true (or "likelihood" probability)

- **Example**:
We examine a group of people. The individual probability $P(C)$ for each member of the group to have cancer is 1 %, i.e., $P(C)$=0.01. This is a "base rate" or "prior" probability (before being inform for the particular case). On the other hand, the probability of being 75 years old is 0.2 %. **What is the probability as a 75 years old to have a cancer, i.e., $P(C|75)$=?**

a) C - event "having cancer" $\Longrightarrow$ probability of having cancer is $P(C) = 0.01$
b) 75 - event "being 75 years old" $\Longrightarrow$ probability to be 75 years old is $P(75) = 0.002$
c) the probability 75 years old to be diagnosed with cancer is 0.5 %, i.e., $P(75|C) = 0.005$

$$P(C|75) = \frac{P(75|C)P(C)}{P(75)} = \frac{0.005 \times 0.01}{0.002} = 2.5\%, \tag{11}$$

(This can be interpreted as: in a sample of 100 000 people, 1000 will have cancer and 200 people will be 75 years old. From these 1000 people - only 5 people will be 75 years old. Thus, of the 200 people, who are 75 years old, only 5 can be expected to have cancer.)

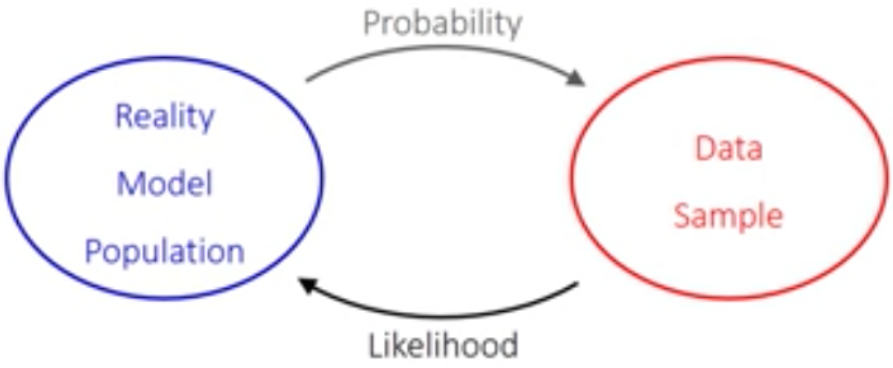

Figure 6: Likelihood function (Reference 14).

5 Maximum Likelihood Function

5.1 General idea

- **Definition**: The likelihood $\mathcal{L}$ of the true parameters being a certain value (given the data) is the same as the probability of observing the data (given some true values; see Fig. 6).

  Given training data set: x_1, x_2, ..., x_n
  Given probability function: P(x_1, x_2, ..., x_n; θ)
  Asked: Find the maximum likelihood estimated of θ

$$\mathcal{L}\infty(\theta|x_1, x_2, ..., x_n) = P(x_1|\theta)\, P(x_2|\theta) \ldots P(x_n|\theta) \tag{12}$$

  Or in short, the likelihood is expressed as the product of the individual probabilities for a given θ,

$$\mathcal{L}(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{m} P(x_i|\theta) \tag{13}$$

- **log-likelihood function**: the maximisation of $\mathcal{L}$ is difficult to calculate as a product of different probabilities; and we find instead the logarithmic function of the likelihood, where this product turns to a sum:

$$\ln \mathcal{L}(\theta|x_1, x_2, ..., x_n) = \sum_{i=1}^{m} \ln P(x_i|\theta) \tag{14}$$

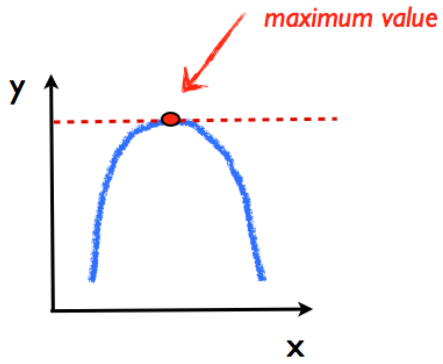

Figure 7: Maximazing log-probability function.

a) maximaze log-probability function
We need to find the maximum value of the log-probability function, corresponding to the optimum best-fit value of the function for a given parameter θ. This is exactly the derivative of the $\ln \mathcal{L}$ with repect to θ made equal to zero:

$$\frac{\partial \ln \mathcal{L}(\theta|x_1, x_2, ..., x_n)}{\partial \theta} = 0 \tag{15}$$

b) verify log-probability function (see Fig. 7)
To find the global maximum of the log-probability function,

$$\frac{\partial^2 \ln \mathcal{L}(\theta|x_1, x_2, ..., x_n)}{\partial \theta} < 0 \tag{16}$$

5.2 Example 1:

Given samples 0,1,0,0,1 from a binomial distribution
Their probabilities are expressed as:
P(x=0)=1-μ
P(x=0)=μ

Requested: What is the maximum likelihood estimated of μ?

Therefore, we define the likelihood as product of the individual probabilities:

$$\begin{aligned}\mathcal{L}(\mu|x_1, x_2, ..., x_n) = P(x=0)\, P(x=1)\, P(x=0)\, P(x=0)\, P(x=1) \;= \\ (1-\mu)\, \mu\, (1-\mu)\, (1-\mu)\, \mu = \mu^2 (1-\mu)^3 \,.\end{aligned} \tag{17}$$

Further, we find the log-likelihood function as

$$\ln \mathcal{L} = 3 \ln(1-\mu) + 2 \ln \mu. \tag{18}$$

We find the maximum value of the log-probability function:

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \tag{19}$$

$$\Rightarrow -3\mu + 2(1-\mu) = 0 \tag{20}$$

$$\Rightarrow \mu = \frac{2}{5} \tag{21}$$

Finally, we verify if we find the global maximum of the log-probability function by the second order derivative:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \mu} = \frac{3(-1)}{(1-\mu)^2} - \frac{2}{\mu^2} \leq 0 \tag{22}$$

Therefore, the value $\mu = \frac{2}{5}$ is indeed our maximum estimate of the log-likelihood function.

5.3 Example 2:

Let's find the likelihood function of data represented by a line in the form $y = f(x) = mx + b$, where any reason for the data to deviate from a linear relation is an added offset in the y-direction. The error y_i was drawn from a Gaussian distribution with a *zero mean* and *known variance* σ_y^2.

In this model, given an independent position x_i, an uncertainty σ_{y_i}, a slope m, an intercept b, the frequence distribution p is:

$$p(y_i|x_i, \sigma_{y_i}, m, b) = \frac{1}{\sqrt{2\pi\sigma_{y_i}^2}} e^{-\frac{|y_i - mx_i - b|^2}{2\sigma_{y_i}^2}}. \tag{23}$$

Therefore, the likelihood will be expressed as:

$$\mathcal{L} = \prod_{i=1}^{N} p(y_i|x_i, \sigma_{y_i}, m, b) \Rightarrow \tag{24}$$

$$\ln \mathcal{L} = K - \sum_{i=1}^{N} \frac{|y_i - mx_i - b|^2}{2\sigma_{y_i}^2} = K - \frac{1}{2}\chi^2, \tag{25}$$

where K is some constant.

Thus, the likelihood maximization is identical to χ^2 minimization !

6 Bayesian generalization

The Bayesian generalization of the frequency distribution p, described in equation 23, have the following expression:

$$p(m, b|\{y_i\}_{i=1}^N, I) = \frac{p(\{y_i\}_{i=1}^N|m, b, I)\, p(m, b|I)}{p(\{y_i\}_{i=1}^N|I)}, \tag{26}$$

where $m, b -$ model parameters
$\{y_i\}_{i=1}^N -$ short-hand for all of the data y_i
$I-$ short-hand for all the prior knowledge of the x_i and σ_{y_i}.

Further, we can read the contributors in equation 26 as the following:
$p(m, b|\{y_i\}_{i=1}^N, I) \rightarrow$ *Posterior* distribution
$p(\{y_i\}_{i=1}^N|m, b, I) \rightarrow$ *Likelihood* $\mathcal{L}$ distribution
$p(m, b|I) \rightarrow$ *Prior* distribution
$p(\{y_i\}_{i=1}^N|I) \rightarrow$ *Normalization* constant

Or,

$$Posterior = \frac{Likelihood \times Prior}{Normalization} \tag{27}$$

7 The Metropolis algorithm

- **First ideas (in the modern time):**
  The algorithm is origially invented by Nicholas Metropolis (1915-1999), but generalized by Wilfred Hastings (1930-2016), later it is called Metropolis-Hastings algorithm.

- **Basic assumptions in the Metropolis algorithm:**
  - assumes a symmetric random walk for the proposed distribution, i.e., $q(x|y) = q(y|x)$

  - the *posterior* distribution is approximated to the Bayesian probabilities since the dominator term in equation 27 is difficult to calculate in practice, but at the same time possible to ignore due to its normalization nature

  - the walkers are keep jumping to explore the parameter space even if they already found the local minimum

  - to improve the efficiency in the MCMC algorithm, the burn-in chains are removed

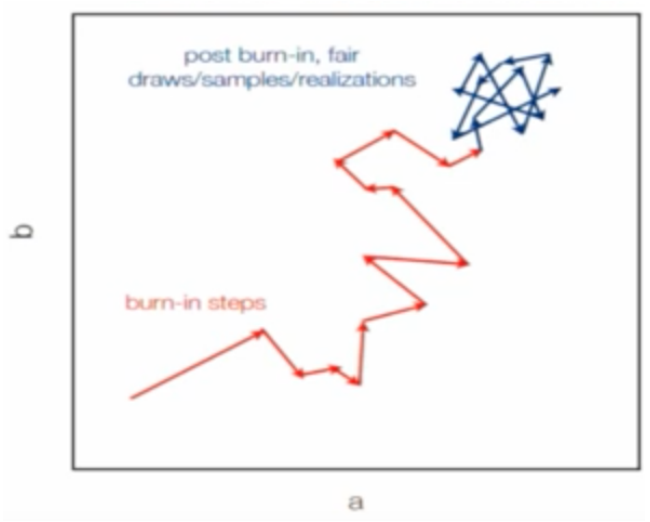

Figure 8: Burn-in and post burn-in steps (credit: Reference 12)

- **Metropolis rules:**

General Case	Ignoring Priors	Ignoring Priors & assuming Gaussian errors
if $P_{trial} > P_i$ accept the jump, so $\theta_{i+1} = \theta_{trial}$	if $\mathcal{L}_{trial} > \mathcal{L}_i$ accept the jump, so $\theta_{i+1} = \theta_{trial}$	$\chi^2_{trial} < \chi^2_i$ accept the jump, so $\theta_{i+1} = \theta_{trial}$
if $P_{trial} < P_i$ accept the jump with probability P_{trial}/P_i	if $\mathcal{L}_{trial} < \mathcal{L}_i$ accept the jump with probability $\mathcal{L}_{trial}/\mathcal{L}_i$	$\chi^2_{trial} > \chi^2_i$ accept the jump with χ^2_{trial}/χ^2_i

where P, $\mathcal{L}$ and χ^2 are the probability, likelihood and chi-squared distributions, repectively. Additionaly, with θ are expressed the positions of the walkers.

References

(1) Kalinova et al., 2017, MNRAS, 464, 1903 (`http://adsabs.harvard.edu/abs/2017MNRAS.464.1903K`)
(2) `http://vcla.stat.ucla.edu/old/MCMC/MCMC_tutorial.htm`
(3) `http://www.mcmchandbook.net/HandbookChapter1.pdf`
(4) `http://physics.ucsc.edu/~drip/133/ch4.pdf`
(5) `https://ned.ipac.caltech.edu/level5/Wall2/Wal3_4.html`
(6) `https://en.wikipedia.org/wiki/Regression_analysis`

(7) https://en.wikipedia.org/wiki/Reduced_chi-squared_statistic
(8) https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo
(9) https://en.wikipedia.org/wiki/Bayes'_theorem
(10) https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
(11) https://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm

Videos:

(12) https://www.youtube.com/watch?v=vTUwEu53uzs
(13) https://www.youtube.com/watch?v=h1NOS_wxgGg
(14) https://www.youtube.com/watch?v=2vh98ful3_M
(15) https://www.youtube.com/watch?v=BFHGIE-nwME
(16) https://www.youtube.com/watch?v=AyBNnkYrSWY