

PATTERN RECOGNITION AND MACHINE LEARNING

Slide Set 2: Estimation Theory

October 2019

Heikki Huttunen
heikki.huttunen@tuni.fi
Signal Processing
Tampere University

Classical Estimation and Detection Theory

- Before the machine learning part, we will take a look at classical estimation theory.
- Estimation theory has many connections to the foundations of modern machine learning.
- Outline of the next few hours:
 - ① Estimation theory:
 - Fundamentals
 - Maximum likelihood
 - Examples
 - ② Detection theory:
 - Fundamentals
 - Error metrics
 - Examples

Introduction - estimation

- Our goal is to estimate the values of a group of parameters from data.
- Examples: radar, sonar, speech, image analysis, biomedicine, communications, control, seismology, etc.
- *Parameter estimation*: Given an N -point data set $\mathbf{x} = \{x[0], x[1], \dots, x[N-1]\}$ which depends on the unknown parameter $\theta \in \mathbb{R}$, we wish to design an *estimator* $g(\cdot)$ for θ

$$\hat{\theta} = g(x[0], x[1], \dots, x[N-1]).$$

- The fundamental questions are:
 - ① What is the model for our data?
 - ② How to determine its parameters?

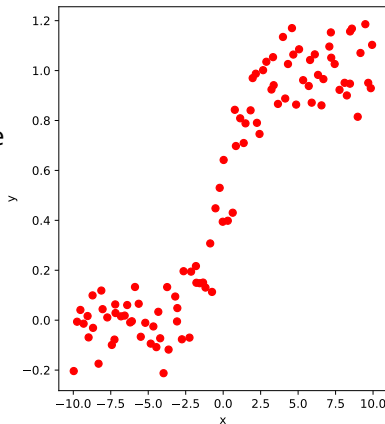
Introductory Example – Straight line

- Suppose we have the illustrated time series and would like to approximate the relationship of the two coordinates.
- The relationship looks linear, so we could assume the following model:

$$y[n] = ax[n] + b + w[n],$$

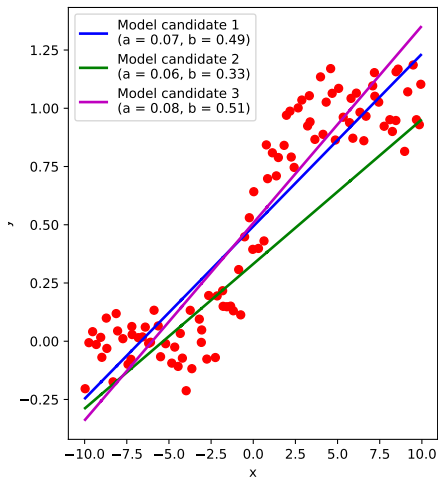
with $a \in \mathbb{R}$ and $b \in \mathbb{R}$ unknown and $w[n] \sim \mathcal{N}(0, \sigma^2)$

- $\mathcal{N}(0, \sigma^2)$ is the normal distribution with mean 0 and variance σ^2 .



Introductory Example – Straight line

- Each pair of a and b represent one line.
- Which line of the three would best describe the data set? Or some other line?



Introductory Example – Straight line

- It can be shown that the best solution (in the *maximum likelihood* sense; to be defined later) is given by

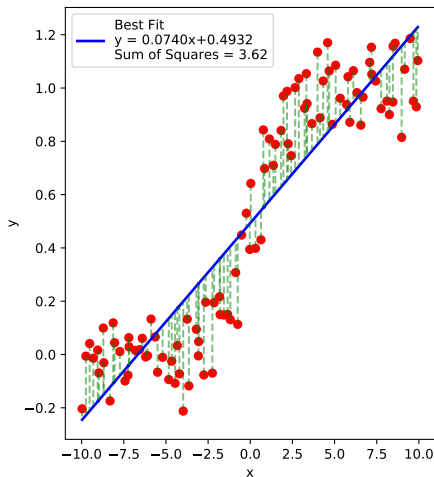
$$\begin{aligned}\hat{a} &= -\frac{6}{N(N+1)} \sum_{n=0}^{N-1} y(n) + \frac{12}{N(N^2-1)} \sum_{n=0}^{N-1} x(n)y(n) \\ \hat{b} &= \frac{2(2N-1)}{N(N+1)} \sum_{n=0}^{N-1} y(n) - \frac{6}{N(N+1)} \sum_{n=0}^{N-1} x(n)y(n).\end{aligned}$$

- Or, as we will later learn, in an easy matrix form:

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

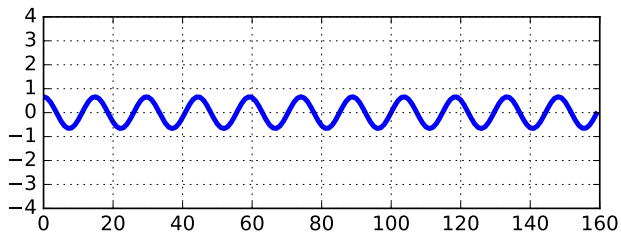
Introductory Example – Straight line

- In this case, $\hat{a} = 0.07401$ and $\hat{b} = 0.49319$, which produces the line shown on the right.
- The line also minimizes the squared distances (green dashed lines) between the model (blue line) and the data (red circles).



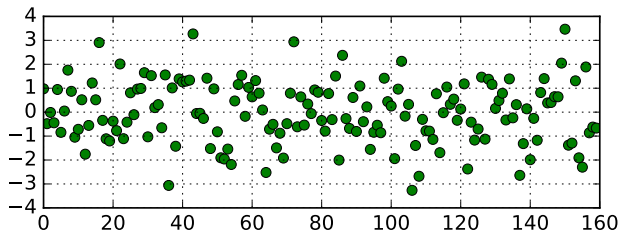
Introductory Example 2 – Sinusoid

- Consider transmitting the sinusoid below.



Introductory Example 2 – Sinusoid

- When the data is received, it is corrupted by noise and the received samples look like below.



- Can we recover the parameters of the sinusoid?

Introductory Example 2 – Sinusoid

- In this case, the problem is to find good values for A , f_0 and ϕ in the following model:

$$x[n] = A \cos(2\pi f_0 n + \phi) + w[n],$$

with $w[n] \sim \mathcal{N}(0, \sigma^2)$.

Introductory Example 2 – Sinusoid

- It can be shown that the *maximum likelihood estimator*; MLE for parameters A , f_0 and ϕ are given by

$$\hat{f}_0 = \text{value of } f \text{ that maximizes } \left| \sum_{n=0}^{N-1} x(n) e^{-2\pi i f n} \right|,$$

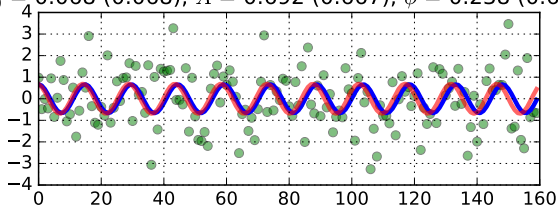
$$\hat{A} = \frac{2}{N} \left| \sum_{n=0}^{N-1} x(n) e^{-2\pi i \hat{f}_0 n} \right|$$

$$\hat{\phi} = \arctan \frac{-\sum_{n=0}^{N-1} x(n) \sin(2\pi \hat{f}_0 n)}{\sum_{n=0}^{N-1} x(n) \cos(2\pi \hat{f}_0 n)}.$$

Introductory Example 2 – Sinusoid

- It turns out that the sinusoidal parameter estimation is very successful:

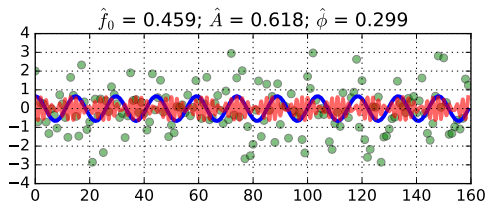
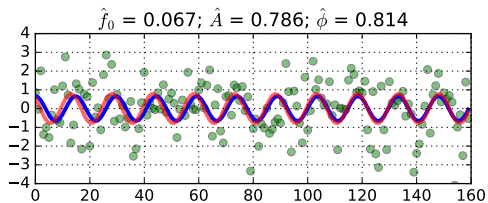
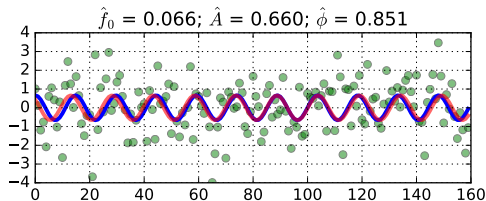
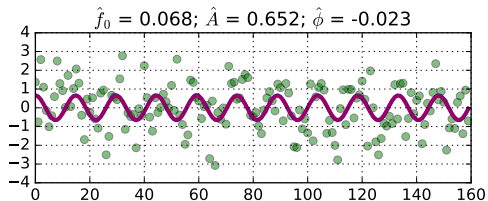
$$\hat{f}_0 = 0.068 \text{ (0.068)}; \hat{A} = 0.692 \text{ (0.667)}; \hat{\phi} = 0.238 \text{ (0.609)}$$



- The blue curve is the original sinusoid, and the red curve is the one estimated from the green circles.
- The estimates are shown in the figure (true values in parentheses).

Introductory Example 2 – Sinusoid

- However, the results are different for each *realization* of the noise $w[n]$.



Introductory Example 2 – Sinusoid

- Thus, we're not very interested in an individual case, but rather on the distributions of estimates
 - What are the expectations: $E[\hat{f}_0]$, $E[\hat{\phi}]$ and $E[\hat{A}]$?
 - What are their respective variances?
 - Could there be a better formula that would yield smaller variance?
 - If yes, how to discover the better estimators?

LEAST SQUARES

Least Squares

- The general solution for linear case is easy to remember.
- Consider the line fitting case: $y[n] = ax[n] + b + w[n]$.
- This can be written in matrix form as follows:

$$\underbrace{\begin{pmatrix} y[0] \\ y[1] \\ \vdots \\ y[N-1] \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} x[0] & 1 \\ x[1] & 1 \\ x[2] & 1 \\ \vdots & \vdots \\ x[N-1] & 1 \end{pmatrix}}_{\mathbf{x}} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\boldsymbol{\theta}} + \mathbf{w}.$$

Least Squares

- Now the model is written compactly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{w}$$

- Solution: The value of $\boldsymbol{\theta}$ minimizing the error $\mathbf{w}^T \mathbf{w}$ is given by:

$$\hat{\boldsymbol{\theta}}_{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- See sample Python code here: https://github.com/mahehu/SGN-41007/blob/master/code/Least_Squares.ipynb

LS Example with two variables

- Consider an example, where microscope images have an uneven illumination.
- The illumination pattern appears to have the highest brightness at the center.
- Let's try to fit a paraboloid to remove the effect:

$$z(x,y) = c_1x^2 + c_2y^2 + c_3xy + c_4x + c_5y + c_6,$$

with $z(x,y)$ the brightness at pixel (x,y) .



LS Example with two variables

- In matrix form, the model looks like the following:

$$\underbrace{\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}}_{\mathbf{z}} = \underbrace{\begin{pmatrix} x_1^2 & y_1^2 & x_1 y_1 & x_1 & y_1 & 1 \\ x_2^2 & y_2^2 & x_2 y_2 & x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^2 & y_N^2 & x_N y_N & x_N & y_N & 1 \end{pmatrix}}_{\mathbf{H}} \cdot \underbrace{\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{pmatrix}}_{\mathbf{c}} + \boldsymbol{\epsilon}, \quad (1)$$

with z_k the grayscale value at k 'th pixel (x_k, y_k) .

LS Example with two variables

- As a result we get the LS fit by $\hat{\mathbf{c}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{z}$,

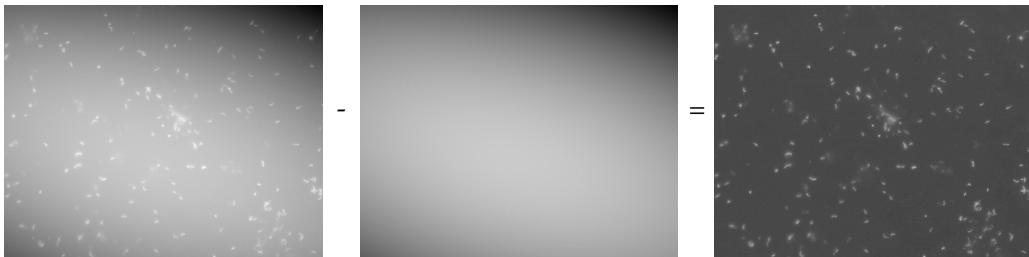
$$\hat{\mathbf{c}} = (-0.000080, -0.000288, 0.000123, 0.022064, 0.284020, 106.538687)$$

- Or, in other words

$$\begin{aligned} z(x,y) = & -0.000080x^2 - 0.000288y^2 \\ & + 0.000123xy + 0.022064x \\ & + 0.284020y + 106.538687. \end{aligned}$$

LS Example with two variables

- Finally, we remove the illumination component by subtraction.



MAXIMUM LIKELIHOOD

Maximum Likelihood Estimation

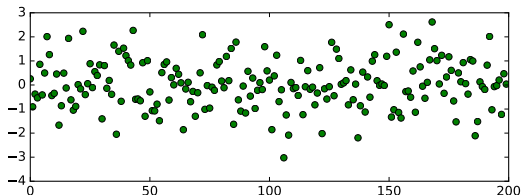
- Maximum likelihood (MLE) is the most popular estimation approach due to its applicability in complicated estimation problems.
- Maximization of likelihood also appears often as the optimality criterion in machine learning.
- The method was proposed by Fisher in 1922, though he published the basic principle already in 1912 as a third year undergraduate.
- The basic principle is simple: find the parameter θ that is the most probable to have generated the data \mathbf{x} .
- The MLE estimator may or may not be optimal in the minimum variance sense. It is not necessarily unbiased, either.

The Likelihood Function

- Consider again the problem of estimating the mean level A of noisy data.
- Assume that the data originates from the following model:

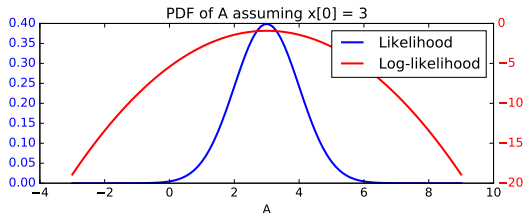
$$x[n] = A + w[n],$$

where $w[n] \sim \mathcal{N}(0, \sigma^2)$: Constant plus Gaussian random noise with zero mean and variance σ^2 .



The Likelihood Function

- The key to maximum likelihood is the *likelihood function*.
- If the probability density function (PDF) of the data is viewed as a function of the unknown parameter (with fixed data), it is called likelihood function.
- Often the likelihood function has an exponential form. Then it's usual to take the natural logarithm to get rid of the exponential. Note that the maximum of the new *log-likelihood* function does not change.



MLE Example

- In our case of estimating the mean of a signal:

$$x[n] = A + w[n], \quad n = 0, 1, \dots, N-1,$$

the likelihood function is easy to construct.

- The noise samples $w[n] \sim \mathcal{N}(0, \sigma^2)$ are assumed independent, so the PDF of the whole batch of samples $\mathbf{x} = (x[0], \dots, x[N-1])$ is obtained by the product rule:

$$p(\mathbf{x}; A) = \prod_{n=0}^{N-1} p(x[n]; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$

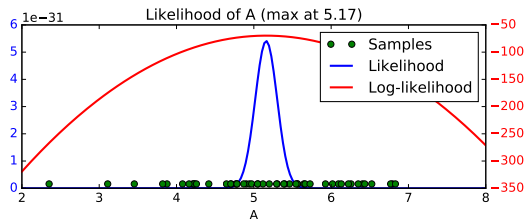
- When we have observed the data \mathbf{x} , we can turn the problem around and consider *what is the most likely parameter A to have generated the data.*

MLE Example

- Some authors emphasize this by turning the order around: $p(A; \mathbf{x})$ or give the function a different name such as $L(A; \mathbf{x})$ or $\ell(A; \mathbf{x})$.
- So, consider $p(\mathbf{x}; A)$ as a function of A and try to maximize it.

MLE Example

- The picture below shows the likelihood function and the log-likelihood function for one possible realization of data.
- The data consists of 50 points, with true $A = 5$.
- The likelihood function gives the probability of observing these particular points with different values of A .



MLE Example

- Instead of finding the maximum from the plot, we wish to have a closed form solution.
- Closed form is faster, more elegant, accurate and numerically more stable.
- Just for the sake of an example, below is the code for the stupid version.

```
# The samples are in array called x0  
  
x = np.linspace(2, 8, 200)  
likelihood = []  
log_likelihood = []  
  
for A in x:  
    likelihood.append(gaussian(x0, A, 1).prod())  
    log_likelihood.append(gaussian_log(x0, A, 1).sum())  
  
print ("Max likelihood is at %.2f" % (x[np.argmax(log_likelihood)]))
```

MLE Example

- Maximization of $p(\mathbf{x}; A)$ directly is nontrivial. Therefore, we take the logarithm, and maximize it instead:

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$
$$\ln p(\mathbf{x}; A) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2$$

- The maximum is found via differentiation:

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)$$

MLE Example

- Setting this equal to zero gives

$$\frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) = 0$$

$$\sum_{n=0}^{N-1} (x[n] - A) = 0$$

$$\sum_{n=0}^{N-1} x[n] - \sum_{n=0}^{N-1} A = 0$$

$$\sum_{n=0}^{N-1} x[n] - NA = 0$$

$$\sum_{n=0}^{N-1} x[n] = NA$$

$$A = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

Conclusion

- *What did we actually do?*
 - We proved that the **sample mean** is the maximum likelihood estimator for the **distribution mean**.
- *But I could have guessed this result from the beginning. What's the point?*
 - We can do the same thing for cases where you can not guess.

Example: Sinusoidal Parameter Estimation

- Consider the model

$$x[n] = A \cos(2\pi f_0 n + \phi) + w[n]$$

with $w[n] \sim \mathcal{N}(0, \sigma^2)$. It is possible to find the MLE for all three parameters:
 $\theta = [A, f_0, \phi]^T$.

- The PDF is given as

$$p(\mathbf{x}; \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \underbrace{(x[n] - A \cos(2\pi f_0 n + \phi))^2}_{w[n]} \right]$$

Example: Sinusoidal Parameter Estimation

- Instead of proceeding directly through the log-likelihood function, we note that the above function is maximized when

$$J(A, f_0, \phi) = \sum_{n=0}^{N-1} (x[n] - A \cos(2\pi f_0 n + \phi))^2$$

is minimized.

- The minimum of this function can be found although it is a nontrivial task (about 10 slides).
- We skip the derivation, but for details, see Kay *et al.* "Statistical Signal Processing: Estimation Theory," 1993.

Sinusoidal Parameter Estimation

- The MLE of frequency f_0 is obtained by maximizing the *periodogram* over f_0 :

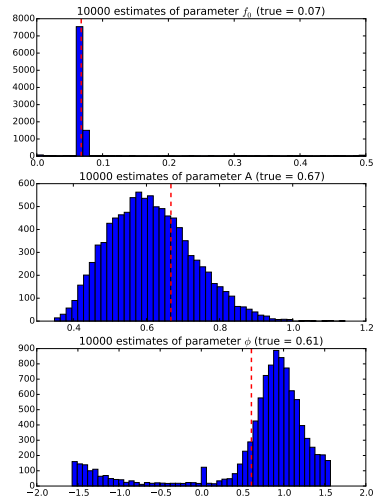
$$\hat{f}_0 = \arg \max_f \left| \sum_{n=0}^{N-1} x[n] \exp(-j2\pi f n) \right|$$

- Once \hat{f}_0 is available, proceed by calculating the other parameters:

$$\hat{A} = \frac{2}{N} \left| \sum_{n=0}^{N-1} x[n] \exp(-j2\pi \hat{f}_0 n) \right|$$
$$\hat{\phi} = \arctan \left(- \frac{\sum_{n=0}^{N-1} x[n] \sin 2\pi \hat{f}_0 n}{\sum_{n=0}^{N-1} x[n] \cos 2\pi \hat{f}_0 n} \right)$$

Sinusoidal Parameter Estimation—Experiments

- Four example runs of the estimation algorithm are illustrated in the figures.
- The algorithm was also tested for 10000 realizations of a sinusoid with fixed θ and $N = 160$, $\sigma^2 = 1.2$.
- Note that the estimator is not unbiased.



Estimation Theory—Summary

- We have seen a brief overview of estimation theory with particular focus on Maximum Likelihood.
- If your problem is simple enough to be modeled by an equation, the estimation theory is the answer.
 - Estimating the frequency of a sinusoid is completely solved by classical theory.
 - Estimating the age of the person in picture can not possibly be modeled this simply and classical theory has no answer.
- **Model based** estimation is the best answer when a model exists.
- Machine learning can be understood as **a data driven approach**.