Pattern Recognition and Machine Learning

Slide Set 2: Estimation Theory

Heikki Huttunen
heikki.huttunen@tut.fi

Department of Signal Processing
Tampere University of Technology

January 2017

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 $\mathbf{x} = \{x[0], x[1], \ldots, x[N-1]\}$ which depends on the unknown parameter $\theta \in \mathbb{R}$, we wish to design an *estimator* $g(\cdot)$ for θ

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

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

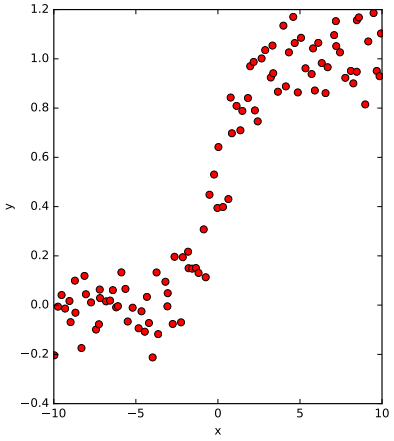
Introductory Example – Straight line

- 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:

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

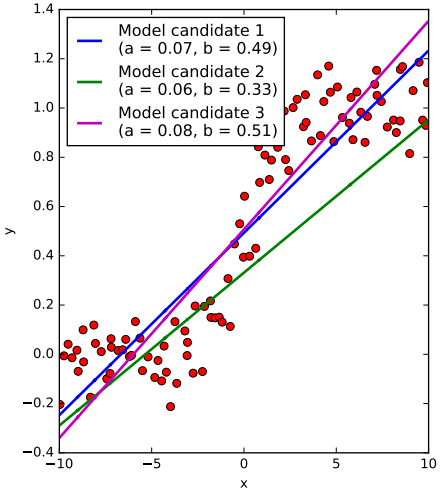
with $a \in \mathbb{R}$ and $b \in \mathbb{R}$ unknown and $w[n] \sim \mathcal{N}(0, \sigma^2)$

- $\mathcal{N}(0, \sigma^2)$ 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

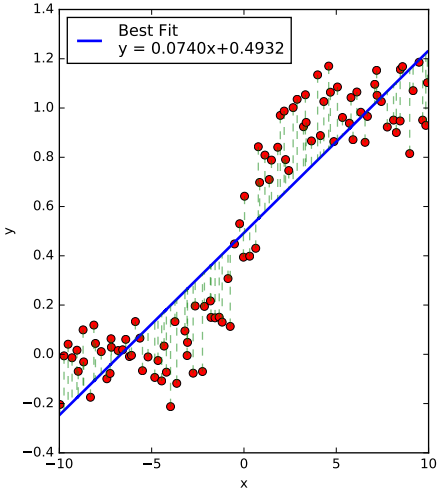
$$\begin{aligned}\hat{a} &= -\frac{6}{N(N+1)}\sum_{n=0}^{N-1} y(n) + \frac{12}{N(N^2-1)}\sum_{n=0}^{N-1} x(n)y(n) \\ \hat{b} &= \frac{2(2N-1)}{N(N+1)}\sum_{n=0}^{N-1} y(n) - \frac{6}{N(N+1)}\sum_{n=0}^{N-1} x(n)y(n).\end{aligned}$$

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

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

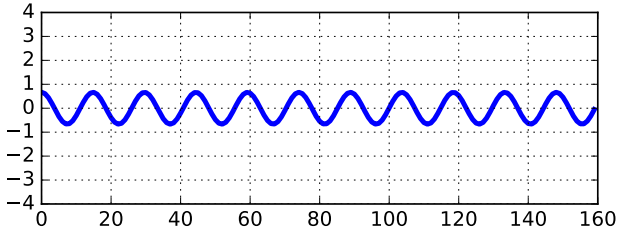
Introductory Example – Straight line

- 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).



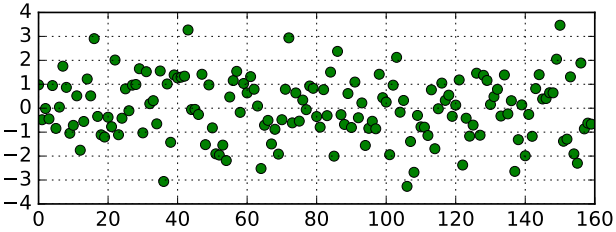
Introductory Example 2 – Sinusoid

- 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

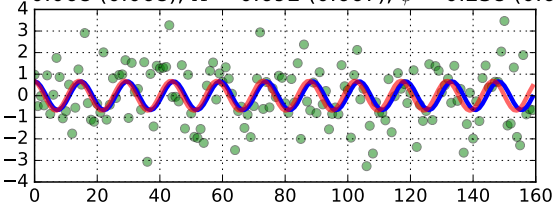
- We will learn that the *maximum likelihood estimator; MLE* for parameters A, f_0 and ϕ are given by

$$\begin{aligned}
\hat{f}_0 &= \text{value of } f \text{ that maximizes } \left| \sum_{n=0}^{N-1} x(n) e^{-2\pi i f n} \right|, \\
\hat{A} &= \frac{2}{N} \left| \sum_{n=0}^{N-1} x(n) e^{-2\pi i \hat{f}_0 n} \right| \\
\hat{\phi} &= \arctan \frac{-\sum_{n=0}^{N-1} x(n) \sin(2\pi \hat{f}_0 n)}{\sum_{n=0}^{N-1} x(n) \cos(2\pi \hat{f}_0 n)}.
\end{aligned}$$

Introductory Example 2 – Sinusoid

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

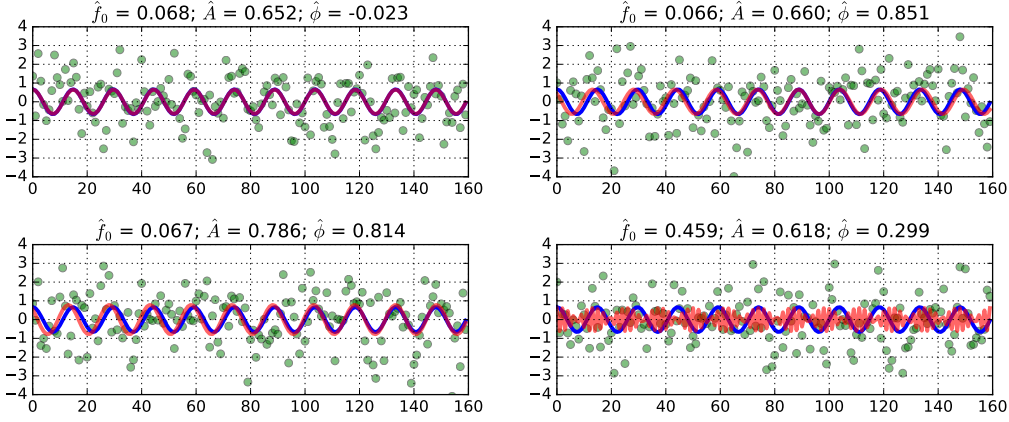
$\hat{f}_0$ = 0.068 (0.068); $\hat{A}$ = 0.692 (0.667); $\hat{\phi}$ = 0.238 (0.609)



- 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 $w[n]$.

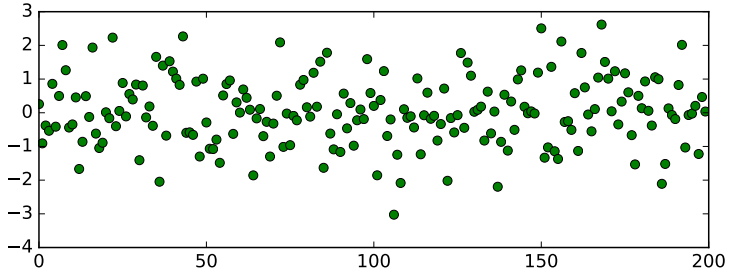


Introductory Example 2 – Sinusoid

- 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?

Example of the Variance of an Estimator

- Consider the estimation of the mean of the following measurement data:



Example of the Variance of an Estimator

- Now we're searching for the estimator $\hat{A}$ in the model

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

  with $w[n] \sim \mathcal{N}(0, \sigma^2)$ where σ^2 is also unknown.
- A natural estimator of A is the sample mean:

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

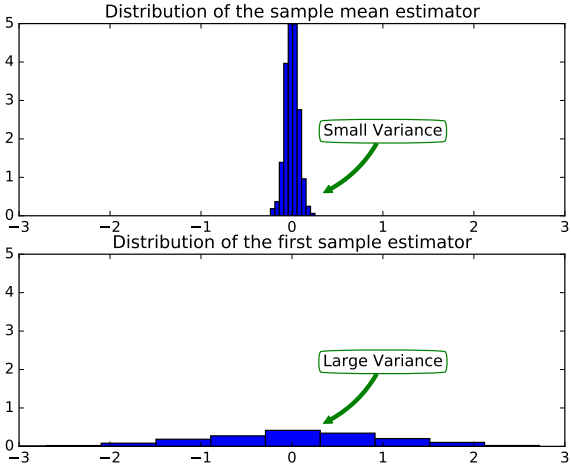
- Alternatively, one might propose to use only the first sample as such:

$$\check{A} = x[0].$$

- How to justify that the first one is better?

Example of the Variance of an Estimator

- Method 1: estimate variances **empirically**.
- Histograms of the estimates over 1000 data realizations are shown on the right.
- In other words, we synthesized 1000 versions of the data with the same statistics.
- Each synthetic sample produces one estimate of the mean for both estimators.
- Code available at https://github.com/mahehu/SGN-41007/blob/master/code/Two_Estimators.ipynb



Example of the Variance of an Estimator

- Method 2: estimate variances **analytically**.
- Namely, it is easy to compute variances in a closed form:

$$\textbf{Estimator 1: } \operatorname{var}(\hat{A}) = \operatorname{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right)$$
$$= \frac{1}{N^2}\sum_{n=0}^{N-1} \operatorname{var}(x[n])$$
$$= \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}.$$
$$\textbf{Estimator 2: } \operatorname{var}(\check{A}) = \operatorname{var}(x[0]) = \sigma^2.$$

Example of the Variance of an Estimator

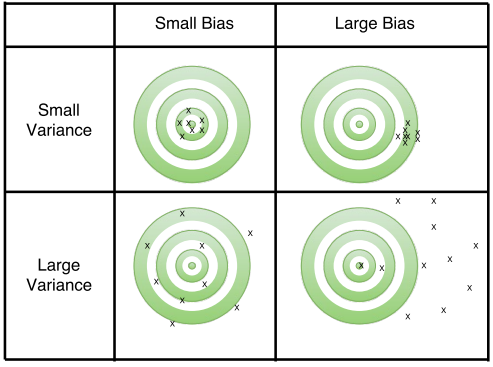
- Compared to the "First sample estimator" $\check{A} = x[0]$, the estimator variance of $\hat{A}$ is one N'th.
- The analytical approach is clearly the desired one whenever possible:
  - Faster, more elegant and less prone to random effects.
  - Often also provides proof that there exists no estimator that would be more efficient.
- Usually can be done for easy cases.
- More complicated scenarios can only be studied empirically.

Estimator Design

- There are a few well established approaches for estimator design:
  - **Minimum Variance Unbiased Estimator (MVU):** Analytically discover the estimator that minimizes the output variance among all *unbiased* estimators.
  - **Maximum Likelihood Estimator (ML):** Analytically discover the estimator that maximizes the likelihood of observing the measured data.
  - Others: **Method of Moments (MoM)** and **Least Squares (LS)**.
- Our emphasis will be on Maximum Likelihood, as it appears in the machine learning part as well.
- Note, that different methods often (not always) result in the same estimator.
- For example, the MVU, ML, MoM and LS estimators for the mean parameter all end up at the same formula: $\hat{A} = \frac{1}{N} \sum x_n$.

Minimum Variance Unbiased Estimator

- Commonly the MVU estimator is considered optimal.
- However, finding the MVU estimator may be difficult. The MVUE may not even exist.
- We will not concentrate on this estimator design approach. Interested reader may consult, *e.g.*, S. Kay: *Fundamentals of Statistical Signal Processing: Volume 1* (1993).
- For an overview, read Wikipedia articles on *Minimum-variance unbiased estimator* and *Lehmann–Scheffé theorem*.



Maximum Likelihood Estimation

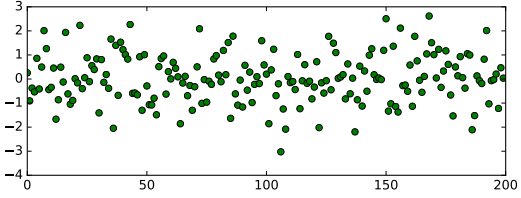
- Maximum likelihood (ML) is the most popular estimation approach due to its applicability in complicated estimation problems.
- Maximization of likelihood also appears often as the optimality criterion in machine learning.
- The method was proposed by Fisher in 1922, though he published the basic principle 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 ML 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

- For simplicity, consider the first sample estimator for estimating A.
- We assume normally distributed $w[n]$, *i.e.,* the following probability density function (PDF):

$$p(w[n]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(w[n])^2\right]$$

- Since $x[n] = A + w[n]$, we can substitute $w[n] = x[n] - A$ above to describe the PDF of $x[n]$[1]:

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

[1]We denote $p(x[n]; A)$ to emphasize that p depends on A.

The Likelihood Function

- Thus, our first sample estimator has the PDF

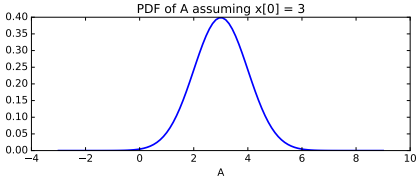
$$p(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0] - A)^2\right]$$

- Now, suppose we have observed $x[0]$, say $x[0] = 3$.
- Then some values of A are more likely than others and we can derive the complete PDF of A easily.

The Likelihood Function

- Actually, the PDF of A has the same form as the PDF of $x[0]$:

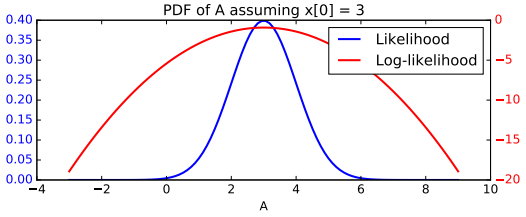
$$\text{pdf of } A = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(3-A)^2\right]$$



- This function is called *the likelihood function* of A, and its maximum the *maximum likelihood estimate*.

The Likelihood Function

- In summary: If the PDF of the data is viewed as a function of the unknown parameter (with fixed data), it is called the *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.



ML Example

- Consider the familiar example of estimating the mean of a signal:

$$x[n] = A + w[n], \qquad n = 0, 1, \ldots, N-1,$$

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

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

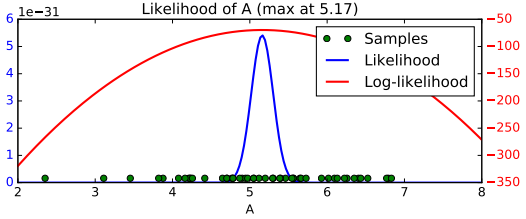
- When we have observed the data $\mathbf{x}$, we can turn the problem around and consider what is the most likely parameter A that generated the data.

ML 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.

ML 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.



ML 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)]))
```

ML Example

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

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

- The maximum is found via differentiation:

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

ML Example

- Setting this equal to zero gives

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

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: $\boldsymbol{\theta} = [A, f_0, \phi]^T$.

- The PDF is given as

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

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:

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

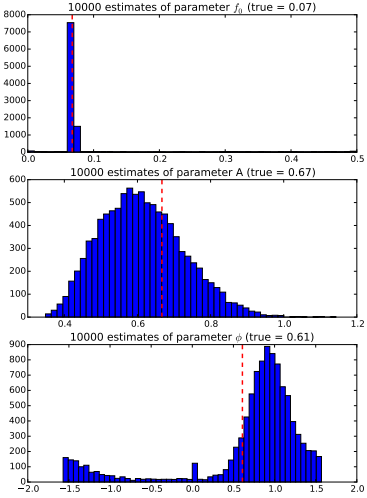
- Once $\hat{f}_0$ is available, proceed by calculating the other parameters:

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

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

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 $\boldsymbol{\theta}$ and $N = 160$, $\sigma^2 = 1.2$.
- Note that the estimator is not unbiased.



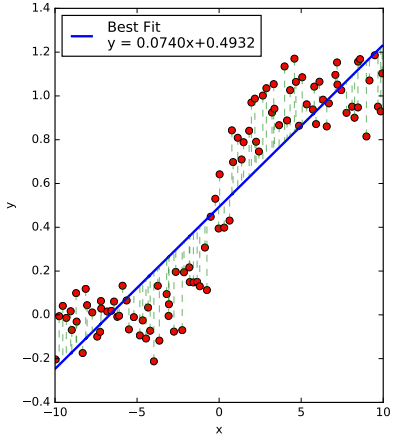
Least Squares

- So far, we have been trying to determine an estimator based on the likelihoods.
- Next we'll consider a class of estimators that in many cases have no optimality properties associated with them, but are still applicable in a number of problems due to their ease of use.
- The *Least Squares* approach originates from 1795, when Gauss invented the method (at the time he was only 18 years old)[2]
- However, the method was properly formalized into its current form and published in 1806 by Legendre.

[2] Compare the date with maximum likelihood from 1912.

Least Squares

- The method became widely known as Gauss was the only one able to describe the orbit of *Ceres*, a minor planet in the asteroid belt between Mars and Jupiter that was discovered in 1801.
- In the least squares approach (LS) we attempt to minimize the squared difference between the given data $x[n]$ and the assumed signal model[3].
- Note, that there is no assumption about the noise PDF.



[3] In Gauss' case, the goal was to minimize the squared difference between the measurements of the location of Ceres and a family of functions describing the orbit.

Least Squares

- The LS estimate $\hat{\theta}_{\text{LS}}$ is the value of θ that minimizes the LS error criterion:

$$J(\theta) = \sum_{n=0}^{N-1} (y[n] - s[n;\theta])^2.$$

- In the line fit case, $\boldsymbol{\theta} = [a, b]^T$ and the LS criterion is

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} (y[n] - s[n;\boldsymbol{\theta}])^2 = \sum_{n=0}^{N-1} (y[n] - (ax[n] + b))^2.$$

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:

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

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:

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

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

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**.