Maximum Likelihood Estimation

- Maximum likelihood (ML) is the most popular estimation approach due to its applicability in complicated estimation problems.
- 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 is in general not optimal in the minimum variance sense. Neither is it unbiased.

Maximum Likelihood Estimation

$\Rightarrow$ If you can choose, take the MVU estimator instead of the ML estimator (if these are different). Generally the MVUE is more difficult to find.

- However, ML estimator is not a poor estimator: *asymptotically* it becomes unbiased and reaches the Cramer-Rao bound.
- Moreover, if an efficient estimator exists, it is the ML estimator.[1]

[1] Remember, an estimator is efficient if it reaches the CRLB. If such an estimator does not exist, we should select the MVUE, which in this case is not necessarily the MLE.

Definition

- The Maximum Likelihood estimate for a scalar parameter θ is defined to be the value that maximizes $p(\mathbf{x};\theta)$.
- $p(\mathbf{x};\theta)$ is the likelihood function and $\ln p(\mathbf{x};\theta)$ is the log-likelihood function. Note that it is equivalent to maximize either of these. Usually the log-likelihood is easier when the distribution is exponential:

$$\begin{aligned}\theta_{ML} &= \arg\max_{\theta}\left(p(\mathbf{x};\theta)\right)\\ &= \arg\max_{\theta}\left(\ln(p(\mathbf{x};\theta))\right)\end{aligned}$$

- In other words, the argument θ that maximizes the function $p(\mathbf{x};\theta)$ or $\ln p(\mathbf{x};\theta)$.

Example

- Consider the familiar example of DC level in WGN:

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

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

- The PDF is

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

- 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.

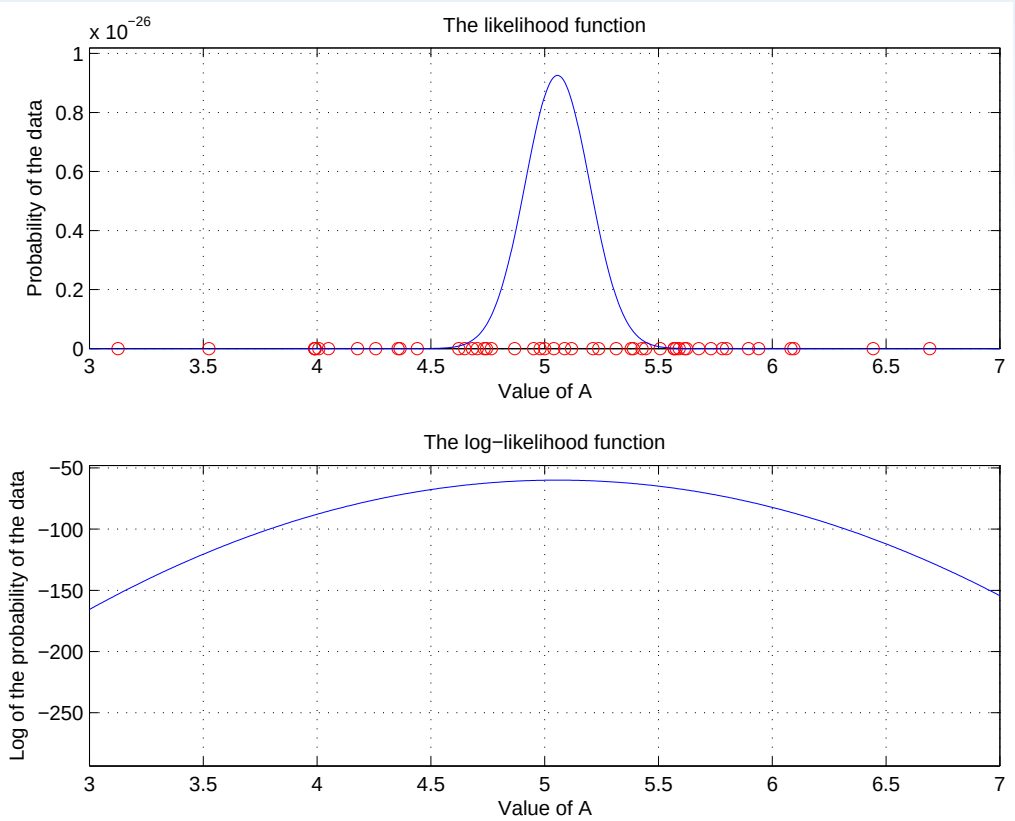
Example

- 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.

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.

Example


The likelihood function
x 10^−26
Probability of the data
Value of A
The log−likelihood function
Log of the probability of the data
Value of A

Example

- Maximization of $p(\mathbf{x}; A)$ is not easy in this case, 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 obtained via differentiation:

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

Example

- Setting this equal to zero gives

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

Example

- Thus, $\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$ is the ML estimator. Since an efficient estimator for this problem exists, the ML estimator reaches it.

Example 2: comparison

- Let us modify the DC level problem such that A is not only the DC level, but also the noise variance:

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

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

- Consider three approaches: **CRLB factorization, RBLS theorem** and **ML estimation**.

Example 2: comparison

1. **CRLB**
   - The CRLB theorem says, that $\hat{A}$ is the MVU (and efficient) estimator if we can factor the the first derivative of the log-likelihood function as follows:

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = I(A)(g(\mathbf{x}) - A)$$

   - In this case the function becomes

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

   - It seems the this function can not be factored into the required form, so CRLB can not be used.

Example 2: comparison

- Note: the CRLB can be shown to be

$$\mathrm{var}(\hat{A}) \geqslant \frac{A^2}{N(A + \frac{1}{2})}.$$

2. RBLS

- The Rao-Blackwell-Lehmann-Scheffe method first searches for a sufficient statistic and then transforms it into an unbiased estimator.
- Can we factor the PDF as

$$p(\mathbf{x}; A) = g(T(\mathbf{x}), \mathbf{A})h(\mathbf{x})$$

- Yes:

Example 2: comparison

$$
\begin{aligned}
p(\mathbf{x};A) &= \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left[-\frac{1}{2A}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\
&= \cdots = \underbrace{\frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\left(\frac{1}{A}\sum_{n=0}^{N-1}x^2[n]+NA\right)\right]}_{g(T(\mathbf{x}),A)} \underbrace{\exp(N\bar{x})}_{h(\mathbf{x})}
\end{aligned}
$$

where $T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$ is the sufficient statistic.

- The next step is to transform $T(\mathbf{x})$ unbiased.
- The expectation is

$$E(T(\mathbf{x})) = \cdots = N(A + A^2)$$

Example 2: comparison

- It seems that there's no way to find a function g such that

$$E(g(T(\mathbf{x}))) = A$$

- The second alternative given by RBLS theorem is not practical, either, because it requires the calculation of

$$E\left(\check{A}|T(\mathbf{x})\right),$$

where $\check{A}$ is any unbiased estimator of A, e.g., $\check{A} = x[0]$.

Example 2: comparison

3. ML

- The log-likelihood function and its derivative has been calculated in the CRLB case:

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

- Let's set it to zero to find the ML estimator:

$$\hat{A}^2 + \hat{A} - \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] = 0.$$

$$\hat{A} = -\frac{1}{2} \pm \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

Example 2: comparison

- Of the two alternatives, we discard the negative one, because variance estimate has to be positive:

$$\hat{A} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}}$$

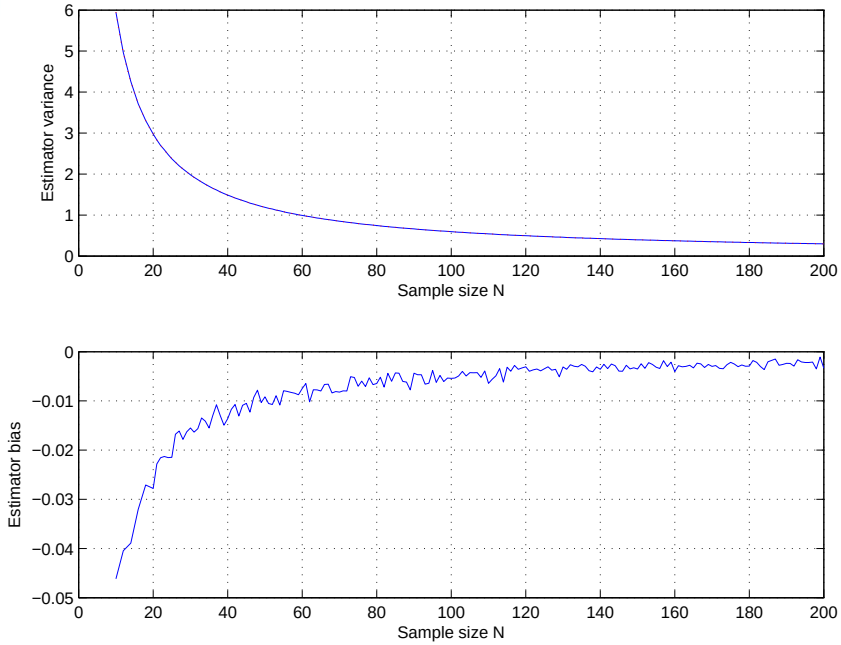
- The estimator is biased since

Example 2: comparison

$$
\begin{aligned}
E(\hat{A}) &= E\left(-\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] + \frac{1}{4}} \right) \\
&\neq -\frac{1}{2} + \sqrt{E\left(\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right) + \frac{1}{4}} \qquad \left(\begin{array}{c}\text{Expectation does not}\\ \text{carry over sqrt}\end{array}\right) \\
&= -\frac{1}{2} + \sqrt{A + A^2 + \frac{1}{4}} \\
&= A.
\end{aligned}
$$

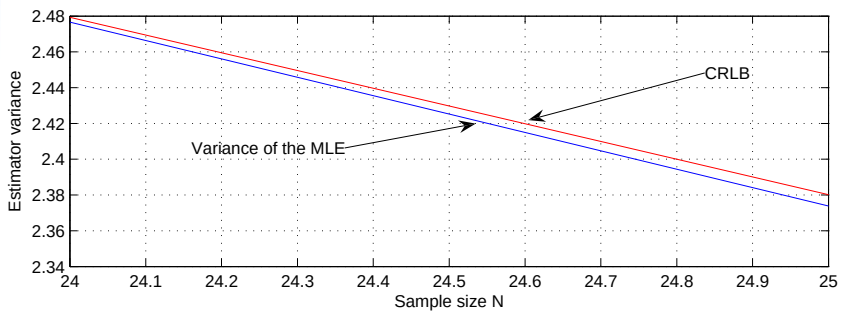
Example 2: comparison

- In practice the bias is negligible when N is large enough: the figure below shows the variance and the bias of the ML estimator as a function of N. The curves are averages of 1000000 realizations.



Example 2: comparison

- For reference, the CRLB is plotted on top of the above figure in red. The MLE has a variance very close to the bound. Zooming in gives the picture below.



- The MLE has actually *smaller* variance than the theoretical limit. This is because the limit only applies to *unbiased* estimators and ours is biased.
- The bias seems to approach zero as $N \to \infty$. This can be shown theoretically:

Example 2: comparison

- By the law of large numbers

$$\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] \to E(x^2[n]) = A + A^2, \text{ when } N \to \infty$$

- Therefore,

$$\hat{A} \to A \text{ (consistent estimator)}$$

- By linearizing we obtain for large N:

$$\hat{A} \approx A + \frac{\frac{1}{2}}{A + \frac{1}{2}} \left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] - (A + A^2) \right].$$

Example 2: comparison

- And therefore,

$$
\begin{aligned}
E(\hat{A}) &\approx E\left(A + \frac{\frac{1}{2}}{A+\frac{1}{2}}\left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] - (A + A^2)\right]\right) \\
&= A + \frac{\frac{1}{2}}{A+\frac{1}{2}}\left[\frac{1}{N}\sum_{n=0}^{N-1} E(x^2[n]) - (A + A^2)\right] \\
&= A + \frac{\frac{1}{2}}{A+\frac{1}{2}}\left[A + A^2 - (A + A^2)\right] = A + 0 = A
\end{aligned}
$$

Example 2: comparison

- Similarly, the linearization gives the asymptotic variance as

$$
\begin{aligned}
\mathrm{var}(\hat{A}) &= \mathrm{var}\left[A + \frac{\frac{1}{2}}{A+\frac{1}{2}}\left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n] - (A + A^2)\right]\right] \\
&= \left(\frac{\frac{1}{2}}{A+\frac{1}{2}}\right)^2 \mathrm{var}\left[\frac{1}{N}\sum_{n=0}^{N-1} x^2[n]\right] \\
&= \frac{\frac{1}{4}}{N(A+\frac{1}{2})^2}\mathrm{var}(x^2[n]) \\
&= \frac{\frac{1}{4}}{N(A+\frac{1}{2})^2} 4A^2(A+\frac{1}{2}) \\
&= \frac{A^2}{N(A+\frac{1}{2})}
\end{aligned}
$$

Example 2: comparison

- As we saw earlier, this is the CRLB.
- Thus, the ML estimator is asymptotically efficient.
- Furthermore, it has a Gaussian pdf.

Summary of properties

- The MLE always becomes optimal and unbiased as $N \to \infty$.
- Sometimes the MLE is optimal with finite sample sizes, as well. More specifically, if an efficient estimator exists (reaching the CRLB), the MLE will be it. If it does not exist, there might be a better alternative than the MLE.
- The asymptotic efficiency and unbiasedness are described in more exact terms in the following theorem.

Summary of properties

Theorem

Asymptotic properties of the MLE: *If the pdf* $p(\mathbf{x};\theta)$ *of the data* $\mathbf{x}$ *satisfies some "regularity" conditions, then the MLE of the unknown parameter* θ *is asymptotically distributed (for large data records) according to*

$$\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, I^{-1}(\theta)) \tag{1}$$

where $I(\theta)$ *is the Fisher information evaluated at the true value of the unknown parameter.*

The "regularity" conditions require the existence of the derivatives of the log-likelihood function, as well as the Fisher information being non-zero.

Example: Sinusoidal parameter estimation

- The main advantage of the MLE is that it is often easy to find in practical problems.
- Consider the phase estimation problem:

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

where A and f_0 are known, $w[n] \sim \mathcal{N}(0, \sigma^2)$ and we are to estimate the phase ϕ.

- The sufficient statistic won't help us, because there's no single statistic.

Example: Sinusoidal parameter estimation

- Instead, in Chapter 5, the book shows that the statistics

$$T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)$$

$$T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$$

  are *jointly sufficient*. However, this does not help us in finding the MVU.
- Let's try the MLE.

Example: Sinusoidal parameter estimation

- The MLE is found by maximizing the likelihood function

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

- The log-likelihood function is now

$$\ln p(\mathbf{x};\phi) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - A\cos(2\pi f_0 n + \phi))^2$$

Example: Sinusoidal parameter estimation

- Differentiating it produces

$$\frac{\partial \ln p(\mathbf{x};\phi)}{\partial \phi} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A\cos(2\pi f_0 n+\phi))A\sin(2\pi f_0 n+\phi)$$

- Setting the differential equal to zero gives an equation for the estimator $\hat{\phi}$:

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

Example: Sinusoidal parameter estimation

After simplification this becomes

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

- Solving $\hat{\phi}$ from this equation is not easy. Therefore we have to resort to the following approximation on the right hand side:

$$\frac{1}{N}\sum_{n=0}^{N-1} \cos(2\pi f_0 n+\hat{\phi})\sin(2\pi f_0 n+\hat{\phi}) = \frac{1}{2N}\sum_{n=0}^{N-1} A\sin(4\pi f_0 n+2\hat{\phi}) \approx 0.$$

Example: Sinusoidal parameter estimation

- This means that even after multiplication by N the term on the right hand side is close to zero. Thus,[2]

$$\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\phi}) \approx 0$$

[2] The effect of this approximation was studied in the mandatory Matlab assignment of the course of 2009. It was discovered that the effect is in fact negligible.

Example: Sinusoidal parameter estimation

- We can solve $\hat{\phi}$ from here using the triangular equality

$$\sin(\alpha + \beta) = \cos\alpha \sin\beta + \sin\alpha \cos\beta$$

or in our case:

$$\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n + \hat{\phi})$$

$$= \sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n) \cos(\hat{\phi}) + \sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n) \sin(\hat{\phi}) \approx 0$$

Example: Sinusoidal parameter estimation

- Further manipulation gives

$$\sin(\hat{\phi}) \sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n) \approx -\cos(\hat{\phi}) \sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n)$$

and

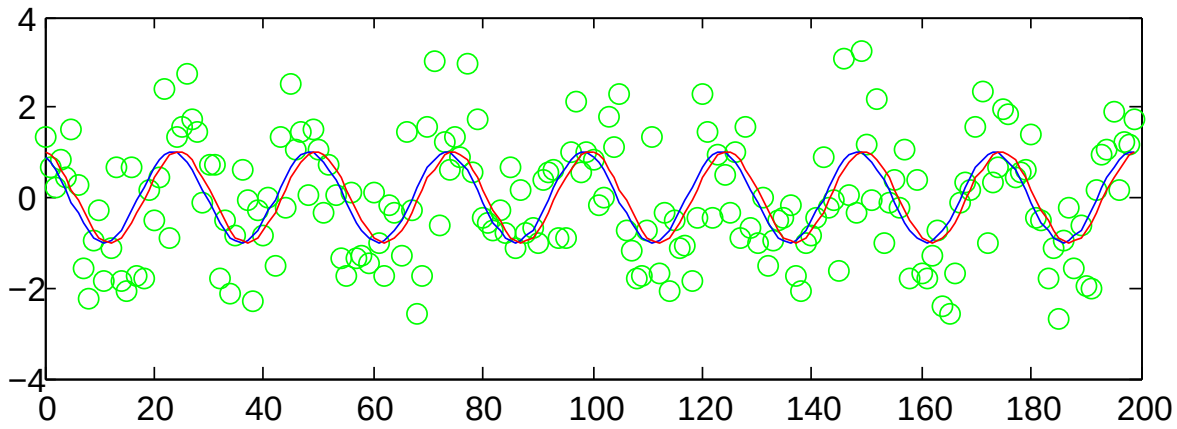
$$\frac{\sin(\hat{\phi})}{\cos(\hat{\phi})} \approx -\frac{\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n)}$$

- Finally, the approximate ML estimator is

$$\hat{\phi} \approx -\arctan \frac{\sum_{n=0}^{N-1} x[n] \sin(2\pi f_0 n)}{\sum_{n=0}^{N-1} x[n] \cos(2\pi f_0 n)}$$

Example: Sinusoidal parameter estimation

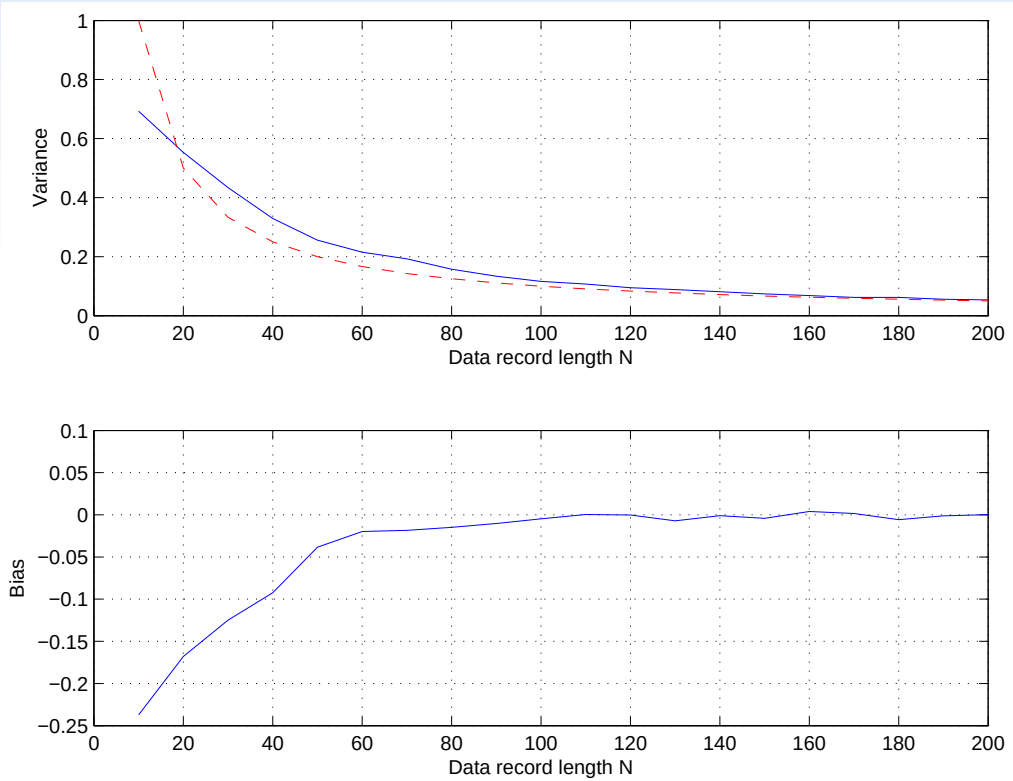
- A result of a simulated test is shown in the picture below. Blue curve shows the true phase, green circles are the noisy data used for the estimation and red line is the estimation result. In this test, $\phi = 0.4$, $N = 100$, $A = 1$, $f_0 = 0.14$ and $\sigma^2 = 5$. The estimate $\hat{\phi} = 0.12$.



Example: Sinusoidal parameter estimation

- The plots below illustrate the performance of the ML estimator as a function of N. For reference, the dashed red curve in the variance plot is the the CRLB, $\mathrm{var}(\hat{\phi}) = \frac{2\sigma^2}{NA^2}$.

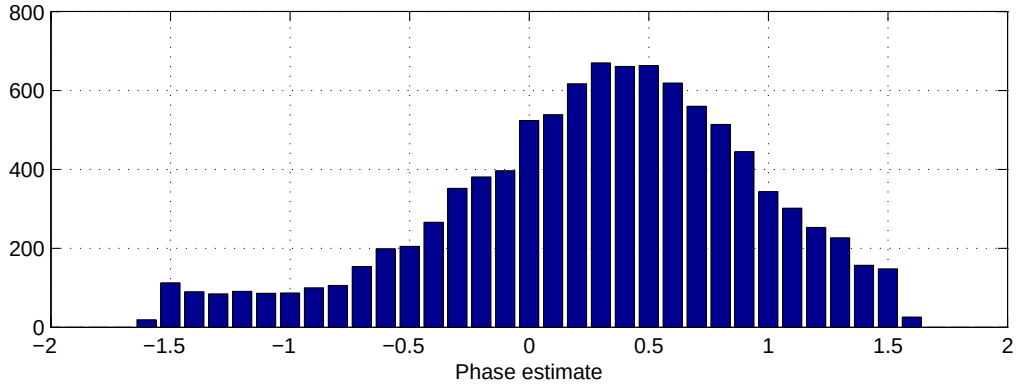
Example: Sinusoidal parameter estimation



- The performance of the estimator seems to get reasonably close to an unbiased optimal estimator after $N \geqslant 100$.

Example: Sinusoidal parameter estimation

- The distribution of the estimates is shown in the below plot for the case $N = 30$. The true value is 0.4.



MLE of transformed parameters

- Often it is required to estimate a transformed parameter instead of the one the PDF depends on.
- For example, in the DC-level problem we might be interested in the power of the signal, A^2 instead of the mean, A.
- Can we transform the ML estimator of A directly: $(\hat{A})^2$?
- For example, consider estimation of transformed DC level in WGN

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

where $w[n]$ is WGN with variance σ^2. We wish to find the MLE of a transformed parameter: $\alpha = \exp(A)$.

MLE of transformed parameters

- The PDF is given by

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

- In terms of the transformed parameter $\alpha = \exp(A)$, the PDF becomes

$$p_T(\mathbf{x}; \alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\ln\alpha)^2\right] \qquad \text{for } \alpha > 0$$

MLE of transformed parameters

- Setting the derivative of $p_T(\mathbf{x};\alpha)$ with respect to α equal to zero yields:

$$\sum_{n=0}^{N-1}(x[n]-\ln\hat{\alpha})\frac{1}{\hat{\alpha}}=0 \quad \Rightarrow \hat{\alpha}=\exp(\bar{x})$$

- Thus, it seems that we can simply transform the estimator.
- Let's see another example before we state the transformation theorem.

MLE of transformed parameters

- Second example: Transformed DC level in WGN with $\alpha = A^2$.
- Now the inverse relation is

$$A = \pm\sqrt{\alpha}$$

because the transformation is not one-to-one.

- If we choose only either $\sqrt{\alpha}$ or $-\sqrt{\alpha}$, then we won't get all possible PDF's. In practise, the estimation would fail if the true A would have a different sign than the square root we chose.

MLE of transformed parameters

- We actually have to consider two pdf's:

$$p_{T_1}(\mathbf{x};\alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\sqrt{\alpha})^2\right] \qquad \text{for } \alpha \geqslant 0$$

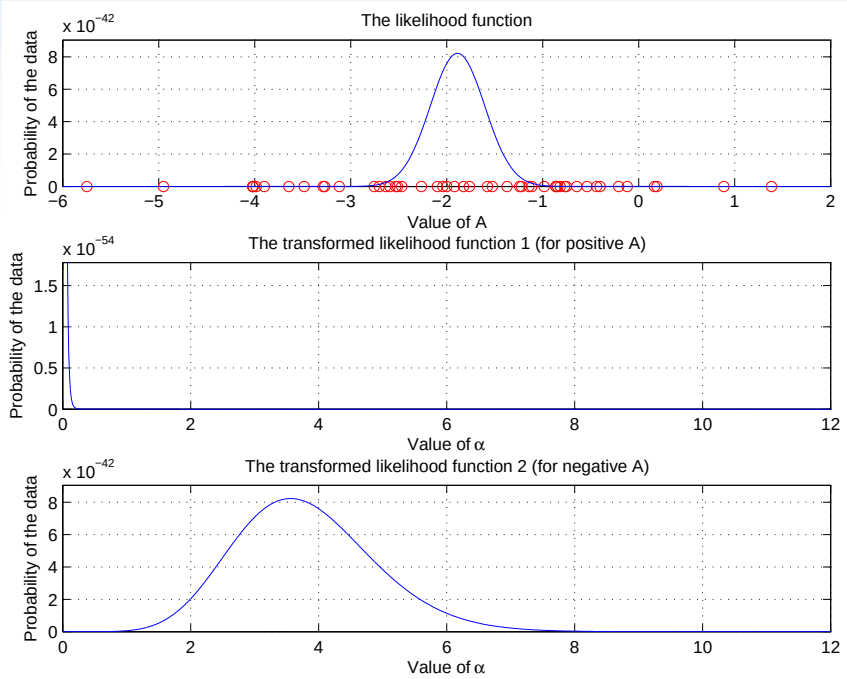
$$p_{T_2}(\mathbf{x};\alpha) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]+\sqrt{\alpha})^2\right] \qquad \text{for } \alpha > 0$$

- Then we'll solve the ML estimation problem in both cases and choose the one that has higher maximum value:

$$\hat{\alpha} = \arg\max_{\alpha}\{p_{T_1}(\mathbf{x};\alpha), p_{T_2}(\mathbf{x};\alpha)\}$$

MLE of transformed parameters

- The three likelihood functions are plotted below.



- In a sense, the first likelihood function is searching for a positive $A = \sqrt{\alpha}$ and the second for negative $A = -\sqrt{\alpha}$.

MLE of transformed parameters

- It is clear that the maximum likelihood estimate for $\alpha = A^2$ is around 3.5 corresponding to $\hat{A} \approx -1.8$. In this example the true $A = -2$.
- It can be shown that the general MLE is $\hat{\alpha} = (\hat{A})^2 = (\bar{x})^2$.

Theorem

Invariance property of the MLE: *The MLE of the parameter* $\alpha = g(\theta)$, *where the pdf* $p(\mathbf{x};\theta)$ *is parameterized by* θ, *is given by*

$$\hat{\alpha} = g(\hat{\theta})$$

where $\hat{\theta}$ *is the MLE of* θ. *The MLE of* θ *is obtained by maximizing* $p(\mathbf{x};\theta)$. *If* g *is a not a one-to-one function, then* $\hat{\alpha}$ *maximizes the modified likelihood function* $\bar{p}(\mathbf{x};\theta)$, *defined as*

$$\bar{p}(\mathbf{x};\alpha) = \max_{\{\theta:\alpha=g(\theta)\}} p(\mathbf{x};\theta).$$