Chapter 8: Least squares (beginning of chapter)
Least Squares

- So far, we have been trying to determine an estimator which was unbiased and had minimum variance.
- 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)\(^1\)
- However, the method was properly formalized into its current form and published in 1806 by Legendre.

\(^1\) Compare the date with maximum likelihood from 1912 and CRLB and RBLS from mid 1900’s.
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\(^2\).

- Note, that there is no assumption about the noise PDF.

---

\(^2\) 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

LS minimizes the total length of vertical distances
Introductory Example

• Consider the model

\[ x[n] = A + Bn + w[n], \quad n = 0, \ldots, N - 1 \]

where \( A \) and \( B \) are unknown deterministic parameters and \( w[n] \sim \mathcal{N}(0, \sigma^2) \).
The LS estimate $\theta_{LS}$ is the value of $\theta$ that minimizes the LS error criterion:

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

where $s[n; \theta]$ is the deterministic part of our model parameterized by $\theta$. 

### Introductory Example
Introductory Example

- In the previous example, $\theta = [A, B]^T$ and the LS criterion is

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n; \theta])^2 = \sum_{n=0}^{N-1} (x[n] - (A + Bn))^2,$$

and the LS estimator is defined as

$$\hat{\theta}_{LS} = \arg \min_{\theta} \sum_{n=0}^{N-1} (x[n] - (A + Bn))^2$$

- Due to the quadratic form, this estimation problem is easily solved, as we will soon see.
The LS problems can be divided into two categories: linear and nonlinear.

The problems where data depends on the parameters in a linear manner are easily solved, whereas nonlinear problems may be impossible to solve in closed form.
Example of a linear LS problem

- Consider estimating the DC level $A$ of a signal:

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

- According to the LS approach, we can estimate $A$ by minimizing

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2$$
Example of a linear LS problem

• Due to the quadratic form, differentiating is easy:

\[
\frac{\partial J(A)}{\partial A} = \sum_{n=0}^{N-1} 2(x[n] - A) \cdot (-1)
\]

• Note that the quadratic form also guarantees that the zero of the derivative is the unique global minimum.
Example of a linear LS problem

- When set equal to zero, we get the minimum of $J(A)$.

\[
N - 1 \sum_{n=0}^{N-1} 2(x[n] - \hat{A}) \cdot (-1) = 0
\]

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

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

\[
\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]
\]
Example of a nonlinear LS problem

- Consider the signal model

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

where the frequency \( f_0 \) is to be estimated. The LSE is defined by the error function

\[ J(f_0) = \sum_{n=0}^{N-1} (x[n] - \cos(2\pi f_0 n))^2 \]
Example of a nonlinear LS problem

• Now, the differentiation gives

\[
\frac{\partial J(f_0)}{\partial f_0} = \sum_{n=0}^{N-1} 2(x[n] - \cos 2\pi f_0 n)(2 \sin 2\pi f_0 n)
\]

\[
= 4 \sum_{n=0}^{N-1} (x[n] \sin 2\pi f_0 n - \cos 2\pi f_0 n \sin 2\pi f_0 n)
\]

• We see that it is not possible to solve the root in closed form.

• The problem could be solved using grid search or iterative minimization.
Example of a nonlinear LS problem

- Non-exhaustive heuristics tend to fail in this kind of problems due to numerous local optima.
- An example of using Matlab’s `nlinfit` is shown below.
- In this example the true frequency was $f_0 = 0.28$. The estimation fails miserably and results in $\hat{f}_0 = 0.02$.
- You can also see that the optimizer gets stuck in a local minimum after a few iterations.
- Note: later we will find an approximate estimator using maximum likelihood principle. This could be used as a starting point for LS, as well.
Linear Least Squares

<table>
<thead>
<tr>
<th>Iteration</th>
<th>SSE</th>
<th>Norm of Gradient</th>
<th>Norm of Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>205.379</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>199.651</td>
<td>162.282</td>
<td>0.000649197</td>
</tr>
<tr>
<td>2</td>
<td>195.271</td>
<td>158.225</td>
<td>0.000578512</td>
</tr>
<tr>
<td>3</td>
<td>192.056</td>
<td>161.282</td>
<td>0.000504393</td>
</tr>
<tr>
<td>4</td>
<td>189.485</td>
<td>139.747</td>
<td>0.000459569</td>
</tr>
<tr>
<td>5</td>
<td>187.202</td>
<td>108.687</td>
<td>0.000441102</td>
</tr>
<tr>
<td>6</td>
<td>185.077</td>
<td>67.3106</td>
<td>0.000433452</td>
</tr>
<tr>
<td>7</td>
<td>183.206</td>
<td>18.3146</td>
<td>0.0004313009</td>
</tr>
<tr>
<td>8</td>
<td>181.814</td>
<td>16.7608</td>
<td>0.000359587</td>
</tr>
<tr>
<td>9</td>
<td>180.980</td>
<td>22.8946</td>
<td>0.000277768</td>
</tr>
<tr>
<td>10</td>
<td>180.587</td>
<td>14.3533</td>
<td>0.000193136</td>
</tr>
<tr>
<td>11</td>
<td>180.418</td>
<td>6.58836</td>
<td>0.000125098</td>
</tr>
<tr>
<td>12</td>
<td>180.352</td>
<td>2.61605</td>
<td>7.77223e-05</td>
</tr>
<tr>
<td>13</td>
<td>180.328</td>
<td>0.970828</td>
<td>4.7173e-05</td>
</tr>
<tr>
<td>14</td>
<td>180.319</td>
<td>0.34898</td>
<td>2.82593e-05</td>
</tr>
<tr>
<td>15</td>
<td>180.316</td>
<td>0.123576</td>
<td>1.6804e-05</td>
</tr>
<tr>
<td>16</td>
<td>180.315</td>
<td>0.0434783</td>
<td>9.94995e-06</td>
</tr>
<tr>
<td>17</td>
<td>180.315</td>
<td>0.0152783</td>
<td>5.87709e-06</td>
</tr>
<tr>
<td>18</td>
<td>180.315</td>
<td>0.00538493</td>
<td>3.46643e-06</td>
</tr>
<tr>
<td>19</td>
<td>180.315</td>
<td>0.00191262</td>
<td>2.04286e-06</td>
</tr>
<tr>
<td>20</td>
<td>180.315</td>
<td>0.000688665</td>
<td>1.20332e-06</td>
</tr>
<tr>
<td>21</td>
<td>180.315</td>
<td>0.000253808</td>
<td>7.08599e-07</td>
</tr>
<tr>
<td>22</td>
<td>180.315</td>
<td>9.67882e-05</td>
<td>4.17201e-07</td>
</tr>
<tr>
<td>23</td>
<td>180.315</td>
<td>3.87321e-05</td>
<td>2.4561e-07</td>
</tr>
</tbody>
</table>

Iterations terminated: relative change in SSE less than OPTIONS.TolFun
Linear Least Squares - Vector Parameter

- We assume that
  \[ s = H\theta \]
  where \( H \) is the *observation matrix*, a known \( N \times p \) matrix \((N > p)\) of full rank \( p \). (No noise pdf assumption!)

- The LSE is found by minimizing
  \[
  J(\theta) = (x - H\theta)^T(x - H\theta)
  = x^T x - x^T H\theta - \theta^T H^T x + \theta^T H^T H \theta
  = x^T x - 2x^T H\theta + \theta^T H^T H \theta
  \]
Linear Least Squares - Vector Parameter

- By setting the gradient \( \frac{\partial J(\theta)}{\partial \theta} = -2H^T x + 2H^T H \theta \) to zero we obtain the normal equations \( H^T H \theta = H^T x \) and the estimator
  \[
  \hat{\theta} = (H^T H)^{-1} H^T x.
  \]

- The minimum LS error
  \[
  J_{\text{min}} = (x - H \theta)^T (x - H \theta) = (x - H(H^T H)^{-1} H^T x)^T (x - H(H^T H)^{-1} H^T x) = x^T (I - H(H^T H)^{-1} H^T) x = x^T (x - H \hat{\theta})
  \]
Idempotence of \((I - H(H^T H)^{-1}H^T)\)

- In the previous slide we used the property that matrix \((I - H(H^T H)^{-1}H^T)\) is idempotent\(^3\).
- Note that the matrix is easy to show to be symmetric.
- The idempotence can be shown as follows:

\[
(I - H(H^T H)^{-1}H^T)(I - H(H^T H)^{-1}H^T) \\
= I - H(H^T H)^{-1}H - H(H^T H)^{-1}H + H(H^T H)^{-1}H^T(H^T H)^{-1}H^T \\
= I - H(H^T H)^{-1}H - H(H^T H)^{-1}H + H(H^T H)^{-1}H^T \\
= I - H(H^T H)^{-1}H
\]

\(^3\) A square matrix \(A\) is idempotent (or projection matrix) if \(AA = A\).
Weighted Linear Least Squares

- The Least squares approach can be extended by introducing weights for the errors of the individual samples.
- This is done by an \( N \times N \) symmetric weighting matrix \( W \):

\[
J(\theta) = (x - H\theta)^T W (x - H\theta)
\]

where \( W \) is a \( N \times N \) positive definite matrix.
- Similar derivation as in the unweighted LS case shows that the weighted LS estimator is given by

\[
\hat{\theta} = (H^T WH)^{-1} H^T W x
\]
Weighted Linear Least Squares

- Note, that setting \( W = I \) gives the unweighted case.
- The minimum LS error is given by

\[
J_{\text{min}} = x^T (W - WH(H^T WH)^{-1} H^T W)x
\]
Weighted LS: example

- Consider the DC level estimation problem

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

where the noise samples have different variances:

\[ w[n] \sim \mathcal{N}(0, \sigma_n^2). \]
Weighted LS: example

- In this case it would be reasonable to choose the weights such that the samples with smaller variance had higher weight, for example,

\[
W = \begin{pmatrix}
\frac{1}{\sigma_0^2} & 0 & 0 & \ldots & 0 \\
0 & \frac{1}{\sigma_1^2} & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \frac{1}{\sigma_{N-1}^2}
\end{pmatrix}
\]
Weighted LS: example

The weighted LS estimator becomes now

$$\hat{A} = (H^TWH)^{-1}H^Twx = \cdots = \frac{\sum_{n=0}^{N-1} x[n]}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$
Chapter 5: General MVU estimation
Previously we saw that it’s possible to find the MVU estimator, if it is efficient (achieves the CRLB).

As a special case, the linear model makes this practical.

In this chapter we’ll consider the case where the MVU does not achieve the bound. It is still possible to find the MVU estimator.

The key to this is so called sufficient statistic, which is a function of the data that contains all available information about it.\(^4\)

---

\(^4\) The Finnish translation is quite descriptive: *tyhjentävä statistiikka*, which says that the statistic drains (exhausts, condenses) all the information from the data.
Let us consider the example of DC level in WGN, where the MVU estimator is

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

In addition to the MVUE, consider the first sample estimator: $\hat{A} = x[0]$

One way of viewing their difference is that the latter one disregards the majority of the samples.
The question arises: which set (or function) of the data would be sufficient to contain all the necessary information?

Possible candidates for a sufficient set of data (sufficient to compute the MVU estimate):

1. \( S_1 = \{x[0], x[1], x[2], \ldots, x[N-1]\} \) (all \( N \) samples as such)
2. \( S_2 = \{x[0] + x[1], x[2], \ldots, x[N-1]\} \) (\( N - 2 \) samples + a sum of two)
3. \( S_3 = \left\{ \sum_{n=0}^{N-1} x[n] \right\} \) (sum of all the samples)
Sufficient statistic

- All of the sets $S_1$, $S_2$ and $S_3$ are sufficient, because the MVUE can be derived from any of them.
- The set $S_3$ is special in the sense that it compresses the information into a single value.
- In general, among all sufficient sets, the one that contains the least amount of elements is called the minimal sufficient statistic. There may be several minimal sufficient statistics: for example in our example it’s easy to come up with other one-element sets of sufficient statistics.
- Next we’ll define sufficiency in mathematical terms.
Definition of Sufficiency

Definition
Consider a set of data $x = \{x[0], x[1], \ldots, x[N - 1]\}$ sampled from a distribution depending on the parameter $\theta$. When estimating $\theta$, a function of the data $T(x)$ is called a sufficient statistic, if the conditional PDF (after observing the value of $T(x)$)

$$p(x \mid T(x) = T_0; \theta)$$

does not depend on the parameter $\theta$.

- A direct use of the above definition is typically difficult. Instead, it is usually more convinient to use Neyman-Fisher factorization theorem that we’ll describe soon.
Definition of Sufficiency

- However, before that, let’s see an example of the difficult way.
Example: Proof of Sufficiency using the Definition

- Consider the familiar DC level in WGN example. Let’s show that the statistic \( T(x) = \sum x[n] \) is sufficient.
- The PDF of the data is
  \[
  p(x; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]
  \]
- We need to show that the conditional PDF
  \( p(x \mid T(x) = T_0; A) \) does not depend on \( A \).
The PDF can be evaluated by the following rule from probability theory:

\[ p(E_1 \mid E_2) = \frac{p(E_1, E_2)}{p(E_2)} \]

In our case this becomes:

\[ p(x \mid T(x) = T_0; A) = \frac{p(x, T(x) = T_0; A)}{p(T(x) = T_0)} \]
Example: Proof of Sufficiency using the Definition

- The denominator \( p(T(x) = T_0) \) is easy, because a sum of Gaussian random variables \( \mathcal{N}(A, \sigma^2) \) is Gaussian with distribution \( \mathcal{N}(NA, N\sigma^2) \), or

\[
p(T(x) = T_0) = p\left( \sum_{n=0}^{N-1} x[n] = T_0 \right)
\]

\[
= \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[ -\frac{1}{2N\sigma^2} (T_0 - NA)^2 \right]
\]

- The numerator \( p(x, T(x) = T_0; A) \) is the joint PDF of the data and the statistic.
Example: Proof of Sufficiency using the Definition

- However, $T(x)$ depends on the data $x$ such that their joint probability takes nonzero values only when $T(x) = T_0$ holds.\(^5\)

- This condition can be written more compactly as:

$$p(x, T(x) = T_0; A) = p(x; A)\delta(T(x) - T_0),$$

where $\delta(\cdot)$ is the *Dirac delta function*\(^6\).

---

\(^5\) This is analogous to the case of two random variables $x$ and $y$ with the relationship $y \equiv 2x$. In this case their joint PDF is $p(x, y) = p(x)\delta(y - 2x)$. The delta term ensures that the joint PDF is zero outside the line $y = 2x$.

\(^6\) The defining characteristic of $\delta(\cdot)$ is

$$\int_{-\infty}^{\infty} p(x)\delta(x) = p(0),$$

i.e., it can be used to pick a single value of a continuous function.
Example: Proof of Sufficiency using the Definition

- Thus, we have

\[
p(x | T(x) = T_0; A) = \frac{p(x, T(x) = T_0; A)}{p(T(x) = T_0)}
\]

\[
= \frac{p(x; A)\delta(T(x) - T_0)}{\frac{1}{\sqrt{2\pi N\sigma^2}} \exp \left[ -\frac{1}{2N\sigma^2} (T_0 - NA)^2 \right]}
\]

\[
= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right] \frac{1}{\sqrt{2\pi N\sigma^2}} \exp \left[ -\frac{1}{2N\sigma^2} (T_0 - NA)^2 \right] \delta(T(x) - T_0)
\]
Example: Proof of Sufficiency using the Definition

- Algebraic manipulation shows that $A$ vanishes from the formula:

$$p(x | T(x) = T_0; A) = \frac{\sqrt{N}}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right] \exp \left[ \frac{T_0^2}{2N\sigma^2} \right] \delta(T(x) - T_0),$$

or, without the delta function:

$$p(x | T(x) = T_0; A) = \begin{cases} \frac{\sqrt{N}}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right] \exp \left[ \frac{T_0^2}{2N\sigma^2} \right], & \text{if } T(x) = T_0, \\ 0, & \text{otherwise.} \end{cases}$$

- Therefore $T(x)$ is sufficient.
Example: Proof of Sufficiency using the Definition

- If nothing else, we learned that using the definition directly is difficult. There's also a more straightforward way, as we'll see next.
Theorem for Finding a Sufficient Statistic

**Theorem**

**Neyman-Fisher Factorization Theorem:**

The function $T(x)$ is a sufficient statistic for $\theta$ if and only if we can factorize the pdf $p(x, \theta)$ as

$$p(x, \theta) = g(T(x), \theta) \ h(x)$$

where $g$ is a function depending on $x$ only through $T(x)$ and $h$ is a function depending only on $x$. 
Example: DC Level in WGN

- Let us find the factorization in the DC level in WGN case. As discussed previously, this should give $T(x) = \sum x[n]$.
- Start with the PDF:

$$p(x; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]$$
Example: DC Level in WGN

- The exponent can be broken into parts:

\[
\sum_{n=0}^{N-1} (x[n] - A)^2 = \sum_{n=0}^{N-1} (x^2[n] - 2Ax[n] + A^2)
\]

\[
= \sum_{n=0}^{N-1} x^2[n] - 2A \sum_{n=0}^{N-1} x[n] + NA^2
\]
Example: DC Level in WGN

- Let’s separate the PDF into two functions: one with $A$ and $T(x)$ only and one with $x$ only:

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

$$g(T(x), A)$$

$$h(x)$$

- Thus,

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

is a sufficient statistic for $A$. 
Example: DC Level in WGN

- Note that the factorization suggests

\[ T(x) = 2 \sum_{n=0}^{N-1} x[n], \]

which is also a sufficient statistic of \( A \).

- In fact, any one-to-one mapping of \( T(x) \) is a sufficient statistic as well, because we can always return to \( T(x) \).
Example: Power of WGN

- Consider estimating the variance $\sigma^2$ of a signal with the following model:

$$x[n] = 0 + w[n],$$

where $w[n] \sim \mathcal{N}(0, \sigma^2)$. Find a sufficient statistic.

- Solution: Start with the PDF:

$$p(x; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right].$$
Example: Power of WGN

• This can be factored in a silly way into the required form:

\[
p(x; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x^2[n] \right] \cdot \frac{1}{g(T(x),\sigma^2)} \cdot h(x).
\]

Thus, \( T(x) = \sum_{n=0}^{N-1} x^2[n] \) is a sufficient statistic for \( \Lambda \).

• Other examples of sufficient statistics follow later as we’re really using the concept to find MVUE.
Finding the MVUE using a sufficient statistic

- Let $T(x)$ be a complete sufficient statistic for a parameter $\theta$. There are two ways of finding an MVU estimator using $T(x)$:
  1. Find any unbiased estimator $\hat{\theta}$ and determine $\hat{\theta} = E(\hat{\theta} \mid T(x))$.
  2. Find some function $g$ so that $\hat{\theta} = g(T(x))$ is an unbiased estimator of $\theta$.

- As one could predict, the latter one is easier to use. That will be the one that we’ll generally employ.

- However, let us find the MVU in our familiar example case (DC level in WGN) using both methods.
Method ①: The difficult way: \( \hat{\theta} = E(\tilde{\theta} \mid T(x)) \)

- First we need a sufficient statistic \( T(x) \) and an unbiased estimator \( \tilde{\theta} \). These can be, for example,

\[
T(x) = \sum_{n=0}^{N-1} x[n] \quad \text{and} \quad \tilde{\theta} = x[0].
\]

- The MVU estimator \( \hat{A} \) can then be expressed in terms of the two as \( \hat{A} = E(\tilde{A} \mid T(x)) \).

- How to calculate?
Method 1: The difficult way: \( \hat{\theta} = E(\tilde{\theta} | T(x)) \)

- In general, the conditional expectation \( E(x | y) \) can be calculated by:

\[
E(x | y) = \int_{-\infty}^{\infty} x \cdot p(x | y) \, dx = \int_{-\infty}^{\infty} x \cdot \frac{p(x, y)}{p(y)} \, dx
\]

- In particular, if \( x \) and \( y \) are Gaussian, this can be shown to be\(^7\)

\[
E(x | y) = E(x) + \frac{\text{cov}(x, y)}{\text{var}(y)}(y - E(y))
\]

\(^7\) See proof in Appendix 10 A of S. Kay’s book or:
http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions
Method 1: The difficult way: \( \hat{\theta} = E(\hat{\theta} \mid T(x)) \)

Therefore, in our case

\[
E(\hat{\theta} \mid T(x)) = \int_{-\infty}^{\infty} \hat{\theta} p(\hat{\theta} \mid T(x)) d\hat{\theta}
\]

\[
= E(\hat{\theta}) + \frac{\text{cov}(\hat{\theta}, T(x))}{\text{var}(T(x))} (T(x) - E(T(x)))
\]

\[
= \mu + \frac{\sigma^2}{N\sigma^2} \left( \sum_{n=0}^{N-1} x[n] - NA \right)
\]

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

which we know is the MVUE.
Method 2: The easy way: Find a Transformation $g$ such that $g(T(x))$ is unbiased

- The easy way:
  We need to find a function $g(\cdot)$ such that $g(T(x))$ is unbiased, i.e.,

  $$E \left( g \left( \sum_{n=0}^{N-1} x[n] \right) \right) = A$$
Method 2: The easy way: Find a Transformation $g$ such that $g(T(x))$ is unbiased

- Because $E(\sum x[n]) = NA$, the function has to satisfy

$$
\frac{E \left( g \left( \sum_{n=0}^{N-1} x[n] \right) \right)}{E \left( \sum_{n=0}^{N-1} x[n] \right)} = \frac{A}{NA} = \frac{1}{N}
$$

- Thus, a simple choice of $g$ is $g(x) = x/N$. Thus

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

- We are now ready to state the theorem named after statisticians Rao and Blackwell. It was extended by Lehmann and Scheffé (the “Additionally...” part); hence the lengthy name.

- The theorem uses the concept of completeness of a statistic that needs to be defined.

- Verification (and even understanding) of completeness can be difficult, so generally we’re just happy to believe that all statistics we’ll see are complete.
The RBLS Theorem

- Definitions of completeness include:
  - A statistic is complete if there is only one function of the statistic that is unbiased.
  - A statistic $T(y)$ is complete if the only real valued function $g$ defined on the range of $T(y)$ that satisfies
    \[ E_\theta(g(T)) = 0, \quad \forall \theta \]
    is the function $g(T) = 0$.
  - You may also check [8] for the definition and examples of completeness.

---

The RBLS Theorem

- For example in our DC level case, the statistic
  \[ T(x) = \sum x[n] \]
  is complete, because the only function for which
  \[ \mathbb{E}[g(T(x))] = 0 \]
  for all \( A \in \mathbb{R} \)
  is the zero function \( g(\cdot) \equiv 0 \).
The RBLS Theorem

Theorem
Rao-Blackwell-Lehmann-Scheffe-theorem
If $\bar{\theta}$ is an unbiased estimator of $\theta$ and $T(x)$ is a sufficient statistic for $\theta$, then $\hat{\theta} = E(\bar{\theta} | T)$ is

1. a valid estimator for $\theta$ (not dependent on $\theta$)
2. unbiased
3. has variance at most that of $\bar{\theta}$, for all $\theta$.

Additionally, if the sufficient statistic is complete, then $\hat{\theta}$ is the MVU estimator.
The RBLS Theorem

Corollary
If the sufficient statistic $T(x)$ is complete, then there’s only one function $g(\cdot)$ that makes it unbiased. Applied to the statistic, $g$ gives the MVUE:

$$\hat{\theta} = g(T(x)).$$

Proof. See Appendix 5B and the discussion on p. 109-110 of Kay’s book.
Example: Mean of Uniform Noise

• Suppose we have observed the data

\[ x[n] = w[n], \]

where \( w[n] \) is i.i.d. noise with PDF \( U(0, \beta) \).\(^9\)

• What is the MVU estimator of the mean \( \theta = \beta / 2 \) from the data \( x \)? Is it just the sample mean as usual?

• Earlier we saw that CRLB regularity conditions do not hold, so that theorem didn’t help us.

• Let’s first find the sufficient statistic using the factorization theorem.

---

\(^9\)The German Tank Problem is an instance of this model: [http://en.wikipedia.org/wiki/German_tank_problem](http://en.wikipedia.org/wiki/German_tank_problem)
Example: Mean of Uniform Noise

- The PDF of the samples can be written as

\[
p(x[n]; \theta) = \frac{1}{\beta} [u(x[n]) - u(x[n] - \beta)],
\]

where \( u(x) \) is the unit step function

\[
u(x) = \begin{cases} 
1, & x \geq 0 \\
0, & x < 0
\end{cases}
\]
Example: Mean of Uniform Noise

• The joint PDF of all the samples is

\[ p(x; \theta) = \frac{1}{\beta^N} \prod_{n=0}^{N-1} [u(x[n]) - u(x[n] - \beta)] \]

• In other words, each value has an equal probability as long as it is in the interval \([0, \beta]\); otherwise zero. This can be written as a formula:

\[ p(x; \theta) = \begin{cases} \frac{1}{\beta^N}, & \text{if } 0 < x[n] < \beta \text{ for all } n = 0, 1, \ldots, N - 1 \\ 0, & \text{otherwise} \end{cases} \]

• This form we cannot factorize as required by the theorem: there are \(N\) conditions that cannot be written as a formula.
Example: Mean of Uniform Noise

• However, we can cleverly express the N conditions using only two: the conditions

\[ 0 < x[n] < \beta \text{ for all } n = 0, 1, \ldots, N - 1 \]

are equivalent to

\[ 0 < \min(x[n]) \text{ and } \max(x[n]) < \beta. \]
Example: Mean of Uniform Noise

Thus,

\[ p(x; \theta) = \begin{cases} \frac{1}{\beta N}, & \text{if } 0 < \min(x[n]) \text{ and } \max(x[n]) < \beta \\ 0, & \text{otherwise} \end{cases} \]

or in a single line using the unit step function \( u(\cdot) \):

\[ p(x; \theta) = \frac{1}{\beta N} u(\beta - \max(x[n])) u(\min(x[n])) \]

\[ g(T(x), \theta) \quad h(x) \]
Example: Mean of Uniform Noise

- Hence we have a factorization required by the Neyman-Fisher factorization theorem, and we know that the sufficient statistic is \( T(x) = \max(x[n]) \).
- Surprise: all the data needed by the MVU is only in the largest sample.
Example: Mean of Uniform Noise

• If we had unlimited time, we could now prove that $T(x)$ is complete. Well, it is.

• In order to find the function $g(\cdot)$ that makes $T(x)$ unbiased we could check what’s $E(T(x))$. We omit this proof as well, and just note that

$$E(T(x)) = \frac{2N}{N + 1} \theta$$

• Therefore, the MVU results from the application of such $g$ that the result is unbiased. Let’s choose

$$g(x) = \frac{N + 1}{2N} x$$
Example: Mean of Uniform Noise

- Result: the MVU of the mean of the uniform distribution is

\[ \hat{\theta} = g(T(x)) = \frac{N + 1}{2N} \max(x[n]). \]
Example: Mean of Uniform Noise

- The optimality can be seen also in practice. The plot below shows the distribution of the MVU estimator (top) in a test of 100 realizations of uniform noise with $\beta = 3$. The other unbiased estimators that one might think of are the sample mean (center) and the sample median (bottom), that both have significantly worse performance than the MVUE.
Example: Mean of Uniform Noise

MVU: Sample variance = 0.00022122. Sample mean = 1.4997

MEAN: Sample variance = 0.0090779. Sample mean = 1.5099

MEDIAN: Sample variance = 0.024855. Sample mean = 1.5179
Both theorems extend to the vector case in a natural way.

**Theorem**

**Neyman-Fisher Factorization - vector parameter:**

If we can factorize the pdf $p(x, \theta)$ as

$$p(x; \theta) = g(T(x), \theta)h(x) \quad (1)$$

where $g$ is a function depending only on $x$ through $T(x) \in \mathbb{R}^{r \times 1}$, and also on $\theta$, and $h$ is a function depending only on $x$, then $T(x)$ is a sufficient statistic for $\theta$. Conversely, if $T(x)$ is a sufficient statistic for $\theta$, then the pdf can be factorized as in (1).
Finding estimator MVU using sufficient statistic - vector parameter

Theorem
Rao-Blackwell-Lehmann-Scheffe - vector parameter
If $\hat{\Theta}$ is an unbiased estimator of $\Theta$ and $T(x)$ is an $r \times 1$ sufficient statistic for $\Theta$, then $\hat{\Theta} = E(\tilde{\Theta} \mid T)$ is

1. a valid estimator for $\Theta$ (not dependent on $\Theta$)
2. unbiased
3. of lesser or equal variance than that of $\tilde{\Theta}$, for all $\Theta$ (each element of $\hat{\Theta}$ has lesser or equal variance).

Additionally, if the sufficient statistic is complete, then $\hat{\Theta}$ is the MVU estimator.
Finding estimator MVU using sufficient statistic - vector parameter

Corollary

If the sufficient statistic $T(x)$ is complete, then there’s only one function $g(\cdot)$ that makes it unbiased. Applied to the statistic, $g$ gives the MVU:

$$\hat{\theta} = g(T(x)).$$
Example: Estimation of Both DC Level and Noise Power

• Consider the DC level in WGN problem,

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

where \( w[n] \) is WGN, and the problem is to estimate \( A \) and the variance \( \sigma^2 \) of \( w[n] \) simultaneously.

• The parameter vector is \( \theta = [A, \sigma^2]^T \).
Example: Estimation of Both DC Level and Noise Power

- The PDF is
  
  \[
  p(x; \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \right]
  \]

- By selecting \( T(x) = [T_1(x), T_2(x)]^T \) with \( T_1(x) = \sum x[n] \) and \( T_2(x) = \sum x^2[n] \), we arrive at the Neyman-Fisher factorization.
Example: Estimation of Both DC Level and Noise Power

- The factorization is now

\[
p(x; \theta) = \frac{1}{(2\pi\sigma^2)^\frac{N}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left( T_2(x) - 2A T_1(x) + NA^2 \right) \right] \cdot \frac{1}{h(x)}
\]

- The sufficient statistic is thus

\[
T(x) = \begin{pmatrix}
\sum_{n=0}^{N-1} x[n] \\
\sum_{n=0}^{N-1} x^2[n]
\end{pmatrix}
\]
Example: Estimation of Both DC Level and Noise Power

• The next step is to find a transformation $g(\cdot)$ that makes the statistic $T(x)$ unbiased.
• Let’s find the expectations:

$$E(T(x)) = \begin{pmatrix}
\sum_{n=0}^{N-1} E(x[n]) \\
\sum_{n=0}^{N-1} E(x^2[n])
\end{pmatrix} = \begin{pmatrix}
NA \\
N(\sigma^2 + A^2)
\end{pmatrix}$$
Example: Estimation of Both DC Level and Noise Power

- The question is: how to remove the bias from both components?
- How about simply dividing by $N$:

$$g(T(x)) = \frac{1}{N} \left( \sum_{n=0}^{N-1} x[n] \right)$$

$$\Rightarrow \mathbb{E}(g(T(x))) = \left( \frac{A}{\sigma^2 + A^2} \right) \neq \left( \frac{A}{\sigma^2} \right)$$

No! The second term won’t get unbiased this easily.
Example: Estimation of Both DC Level and Noise Power

- What about trying to cancel the extra $A^2$ term by using $T_1(x)$:

\[
g(T(x)) = \left( \frac{1}{N} T_1(x) \right) = \left( \frac{1}{N} T_2(x) - \left[ \frac{1}{N} T_1(x) \right]^2 \right) = \left( \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - \bar{x}^2 \right)
\]

- This way the expectation becomes quite close to the true value:

\[
E(g(T(x))) = \left( \frac{1}{N} \sum_{n=0}^{N-1} x^2[n] - \bar{x}^2 \right) = \left( \sigma^2 + A^2 - E(\bar{x}^2) \right)
\]
Example: Estimation of Both DC Level and Noise Power

- Since $\bar{x}$ is Gaussian, we know the expectation of its square: $E(\bar{x}^2) = \Lambda^2 + \sigma^2/N$, which yields

$$E(g(T(x))) = \left(\frac{\Lambda}{\sigma^2 + \Lambda^2 - \Lambda^2 - \sigma^2/N}\right) = \left(\frac{\Lambda}{(N-1)\sigma^2/N}\right)$$

If we multiply the second term with $N/(N-1)$ the result will become unbiased. Thus, the MVUE is

$$\hat{\theta} = \left(\frac{N}{N-1}\right) \left(\frac{1}{N} \sum_{n=0}^{N-1} x[n] - \bar{x}^2\right)$$
Example: Estimation of Both DC Level and Noise Power

- The estimator (or its second term) is not efficient. It can be shown that the variance of the second term is larger than the CRLB:

\[
\mathbf{C}_\hat{\theta} = \begin{pmatrix}
\frac{\sigma^2}{N} & 0 \\
0 & \frac{2\sigma^4}{N-1}
\end{pmatrix}
\quad \text{while} \quad
\mathbf{I}^{-1}(\theta) = \begin{pmatrix}
\frac{\sigma^2}{N} & 0 \\
0 & \frac{2\sigma^4}{N}
\end{pmatrix}
\]