

EE6617: 03 General MVU and Best Linear Unbiased Estimators (BLUE)

A **general approach to design the MVU estimator and best linear unbiased estimator (BLUE)**, which are not necessarily **efficient**, will be discussed in this lecture note.

Major references:

- *Fundamentals of Statistical Processing, Volume I: Estimation Theory*, 1/E, by Steven M. Kay, Prentice Hall, 1993
- *Statistical Signal Processing*, 1/E, by Louis L. Scharf, Prentice Hall, 1991

We know from the previous lecture that the **CRLB** can tell us

- a **performance bound** of the estimator $\hat{\theta}$
- a trial methodology to design the **efficient estimator** (i.e. it may fail to find the efficient estimator).

However, the applicability of CRLB is quite limited, except for the **linear MVU estimator**. Notice that

$$\text{Efficient Estimator} \Rightarrow \text{MVU Estimator},$$

but the **MVU estimator** is not necessarily the **efficient estimator**. The **CRLB** only provides a method to design the **efficient estimator, but not for a general MVU estimator**.

Even if an efficient estimator does not exist, it is still of interest to be able to find the MVU estimator. To achieve this, we first argue the concept of **sufficient statistics**.

1 Sufficient Statistics

It is always noteworthy that an **estimator is a function of the observed data set $\mathbf{x}$, $\theta = f(\mathbf{x})$** .

Suppose the problem of estimating a DC level A given the observation $\mathbf{x} = \mathbf{1}A + \mathbf{w}$. We already know the **efficient** estimator is $\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$ and $\text{var}(\hat{A}) = \frac{\sigma^2}{N}$. If, on the other hand, we had chosen $\hat{A} = x[0]$, it is immediately clear, based on our previous discussion, its variance (being σ^2) is much larger than $\frac{\sigma^2}{N}$.

Q: How can we judge if our data set has or does not have all the information about A in $\mathbf{x}$?

Definition 1 A data set is said to be **sufficient** if it contains all the information pertinent to the estimation problem.

Consider following data sets for the DC level A estimation problem.

$$\begin{aligned}\mathcal{S}_1 &= \{x[0], x[1], \dots, x[N-1]\} \\ \mathcal{S}_2 &= \{x[0] + x[1], x[3], \dots, x[N-1]\} \\ \mathcal{S}_3 &= \left\{ x[0], \sum_{n=1}^{N-1} x[n] \right\} \\ \mathcal{S}_4 &= \left\{ T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] \right\}.\end{aligned}$$

It is clear that $\mathcal{S}_1, \mathcal{S}_3, \mathcal{S}_4$ are sufficient to get $\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$, but $\mathcal{S}_2$ is not. However, once we know $\mathcal{S}_4$, the individual data are not needed and we call $\mathcal{S}_4$ the **minimal sufficient statistics**.

Q: What makes the sufficient statistics?

As mentioned before, the **PDF (or likelihood function)** contains all the necessary information pertinent to the estimation problem. To quantify the sufficient statistics, we first look at our running example. Write the PDF of $\mathbf{x} = \mathbf{1}A + \mathbf{w}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$,

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

Assume that $T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] = T_0$ has been observed, where T_0 is a realization of $T(\mathbf{x})$. This observation will change the PDF to the conditional one $p(\mathbf{x}|T(\mathbf{x}) = T_0; A)$. If **$T(\mathbf{x})$ is sufficient to design an efficient estimator**, this should allow the **conditional $p(\mathbf{x}|T(\mathbf{x}) = T_0; A)$ independent of A , i.e.,**

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = p(\mathbf{x}|T(\mathbf{x}) = T_0). \quad (1)$$

This can be checked as below. By **Bayes rule**,

$$p(\mathbf{x}|T(\mathbf{x}) = T_0; A) = \frac{p(\mathbf{x}, T(\mathbf{x}) = T_0; A)}{p(T(\mathbf{x}) = T_0, A)}$$

and since $T(\mathbf{x}) \sim \mathcal{N}(NA, N\sigma^2)$

$$p(T(\mathbf{x}) = T_0; A) = \frac{1}{(2\pi N\sigma^2)^{1/2}} \exp \left(-\frac{1}{2N\sigma^2} (T_0 - NA)^2 \right).$$

Notice that **the joint distribution $p(\mathbf{x}, T(\mathbf{x}) = T_0; A)$ is only non-zero when $T(\mathbf{x}) = T_0$,**

meaning $p(\mathbf{x}, T(\mathbf{x}) = T_0; A) = p(\mathbf{x}; A) \delta(T(\mathbf{x}) = T_0)$. Now, we check (1),

$$\begin{aligned}
p(\mathbf{x}|T(\mathbf{x}) = T_0; A) &= \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right) \delta(T(\mathbf{x}) = T_0)}{\frac{1}{(2\pi N\sigma^2)^{1/2}} \exp\left(-\frac{1}{2N\sigma^2} (T_0 - NA)^2\right)} \\
&= \frac{\sqrt{N}}{(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp\left(-\frac{\sum_{n=1}^{N-1} x^2[n] - 2AT_0 + NA^2}{2\sigma^2} + \frac{\left(\frac{T_0^2}{N} - 2AT_0 + NA^2\right)}{2\sigma^2}\right) \delta(T(\mathbf{x}) = T_0) \\
&= \frac{\sqrt{N}}{(2\pi\sigma^2)^{\frac{N-1}{2}}} \exp\left(-\frac{\sum_{n=1}^{N-1} x^2[n]}{2\sigma^2} + \frac{\left(\frac{T_0^2}{N}\right)}{2\sigma^2}\right) \delta(T(\mathbf{x}) = T_0)
\end{aligned}$$

$\Rightarrow p(\mathbf{x}|T(\mathbf{x}) = T_0; A)$ does not depend on A . Thus, $T(\mathbf{x})$ is a sufficient statistic of A .

This example illustrates **a one** approach to determine the sufficient statistics. However,

- **evaluating the conditional PDF** $p(\mathbf{x}|T(\mathbf{x}) = T_0, A)$ is not simple
- **identifying potential sufficient statistics** is more complicated.

$\Rightarrow$ we need more efficient methods to identify the sufficient statistic.

2 Neyman-Fisher Factorization

The **Neyman-Fisher Factorization** provides a simple procedure for finding sufficient statistics.

Theorem 1 *If we can factorize $p(\mathbf{x}; \theta)$ as*

$$p(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta) h(\mathbf{x}) \quad (2)$$

where $g(\cdot)$ is a function of $T(\mathbf{x})$ (not a function of $\mathbf{x}$) and $h(\cdot)$ is a function of only $\mathbf{x}$ (with no θ dependency), then $T(\mathbf{x})$ is a sufficient statistic for θ . Conversely, for any sufficient statistic $T(\mathbf{x})$, a factorization (2) exists.

Proof See Section 6.1.

Example 1: DC Level in Noise

We apply Neyman-Fisher theorem to find the sufficient statistics for the observation $\mathbf{x} = \mathbf{1}A + \mathbf{w}$ where

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

The $p(\mathbf{x}; A)$ is

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

Thus, $T(\mathbf{x}) = \sum_{n=1}^{N-1} x[n]$ is a sufficient statistics. ■

Example 2: Noise Power

Assume we observe $\mathbf{x} = \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and want to estimate σ^2 .

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

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

Q: How to use sufficient statistics to get an MVU estimator?

3 Using Sufficient Statistics to Find the MVU Estimator

Provided a sufficient statistic $T(\mathbf{x})$ for θ , we can make use of the **Rao-Blackwell-Lehmann-Scheffe (RBLs) theorem** to find the MVU estimator. To fulfill this, we need an important definition of **complete sufficient statistic**.

Definition 2 The sufficient statistic $T(\mathbf{x})$ is **complete** if there is **only one function** $g(\cdot)$ **such that** $E[g(T(\mathbf{x}))] = \theta$.

Theorem 2 (RBLs) Let $\check{\theta}$ be **any** unbiased estimator of θ and $T(\mathbf{x})$ be a sufficient statistic. Then, the estimator $\hat{\theta} = E[\check{\theta}|T(\mathbf{x})]$ is

- 1) A **valid estimator** of θ (solely a function of $\mathbf{x}$)
- 2) Unbiased

$$3) \quad \text{var}(\hat{\theta}) \leq \text{var}(\check{\theta}).$$

If the sufficient statistic is **complete**, the $\hat{\theta}$ is the **MVU estimator**.

Proof See Section 4.2.

Example 3: Incomplete Sufficient Statistic

Consider a single observation

$$x[0] = A + w[0]$$

where $w[0] \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ with $\mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$ being the uniform distribution over the interval $[-\frac{1}{2}, \frac{1}{2}]$. It is obvious that the sufficient statistic is $T = x[0]$. Furthermore, we may conclude that $g(x[0]) = x[0]$ is a viable candidate for the MVU estimator. However, whether or not $g(x[0]) = x[0]$ is MVU estimator still requires us to verify that $T = x[0]$ is complete statistic.

Claim there exists another function $h(T)$ with the unbiased property $E[h(x[0])] = x[0]$. Notice that if $g = h$, then T is complete and otherwise, it is incomplete. Let $v(T) = g(T) - h(T)$ and examine whether the only possible solution for

$$\int v(T)p(\mathbf{x}; A)d\mathbf{x} = 0 \text{ for all } A$$

is $v(T) = 0$ or not.

Since

$$p(\mathbf{x}; A) = p(T; A) = \begin{cases} 1 & \text{if } A - \frac{1}{2} \leq T \leq A + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

the condition is reduced to

$$\int_{A-\frac{1}{2}}^{A+\frac{1}{2}} v(T)dT = 0 \text{ for all } A.$$

It is straightforward to find the non-zero function $v(T) = \cos(2\pi T)$ that satisfies the condition, i.e., $\int_{A-\frac{1}{2}}^{A+\frac{1}{2}} \cos(2\pi T)dT = 0$. Hence, $v(T) = g(T) - h(T) = T - h(T) = \cos(2\pi T)$ or

$$h(x[0]) = x[0] - \cos(2\pi x[0]).$$

Thus, the estimator $\hat{A} = h(x[0])$ is unbiased.

Since we have found one other unbiased estimator that is also a function of the sufficient statistic, the sufficient statistic is not complete and we can not apply RBLS theorem. Thus, we can not assert $E[x[0]|T = x[0]] = x[0]$ is MVU estimator. ■

Remark 1 We say that a sufficient statistic is complete if the condition

$$\int v(T)p(T; \theta)dT = 0 \text{ for all } \theta \quad (3)$$

is satisfied only by $v(T) = 0$ for all T .

For many practical cases of interest the completeness of sufficient statistic holds. In particular, for the **exponential family of PDFs (Gaussian PDF, exponential PDF, chi-square PDF, etc.)** the completeness satisfies.

The **general MVU estimator design procedure (when we can not apply CRLB)** is

1. Apply **Neyman-Fisher factorization theorem** to find a single sufficient statistic $T(\mathbf{x})$.
2. Determine if $T(\mathbf{x})$ is **complete or not**. If so, proceed; if not, this approach can not be used.
3. Find a function g such that $\hat{\theta} = g(T(\mathbf{x}))$ **yields unbiased estimator**. Then, $\hat{\theta}$ is the MVU estimator.

4 Extension to a Vector Parameter

Many of results in the previous sections can be easily extended to vector case.

Definition 3 A vector statistic $\mathbf{T}(\mathbf{x}) = [T_1(\mathbf{x}) \ T_2(\mathbf{x}) \ \dots \ T_r(\mathbf{x})]^T$ is said to be **sufficient** for the estimation of $\boldsymbol{\theta}$ if $p(\mathbf{x}|\mathbf{T}(\mathbf{x}); \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{T}(\mathbf{x}))$.

Theorem 3 (Neyman-Fisher Factorization for vector parameter) If we can factor the PDF $p(\mathbf{x}; \boldsymbol{\theta})$ as

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

where $\mathbf{T}(\mathbf{x}) = [T_1(\mathbf{x}) \ T_2(\mathbf{x}) \ \dots \ T_r(\mathbf{x})]^T \in \mathbb{R}^{r \times 1}$, then $\mathbf{T}(\mathbf{x})$ is a sufficient statistic for $\boldsymbol{\theta}$, and vice versa.

Theorem 4 If $\check{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ and $\mathbf{T}(\mathbf{x})$ is a sufficient statistic, then $\hat{\boldsymbol{\theta}} = E[\check{\boldsymbol{\theta}}|\mathbf{T}(\mathbf{x})]$ is

- 1) A **valid estimator** of $\boldsymbol{\theta}$ (solely a function of $\mathbf{x}$)
- 2) Unbiased
- 3) $\text{var}\left(\left[\hat{\boldsymbol{\theta}}\right]_i\right) \leq \text{var}\left(\left[\check{\boldsymbol{\theta}}\right]_i\right)$.

If the sufficient statistic is **complete**, the $\hat{\boldsymbol{\theta}}$ is the **MVU estimator**.

Remark 2 We say that a sufficient statistic is complete if the condition

$$\int \mathbf{v}(\mathbf{T})p(\mathbf{T}; \boldsymbol{\theta})d\mathbf{T} = \mathbf{0} \text{ for all } \boldsymbol{\theta}$$

is satisfied only by $\mathbf{v}(\mathbf{T}) = \mathbf{0}$ for all $\mathbf{T}$.

5 Best Linear Unbiased Estimator (BLUE)

It would occur often in practice that we may **not know the PDF of the data**. In this case, **our previous methods, relying on CRLB, Neyman-Fisher, and RBLs can not be applied**. Even if the PDF is known, **CRLB** and **RBLs** theorems do **not guarantee** to produce the MVU estimator.

In many of cases, it is reasonable to resort to a **suboptimal estimator**. A common approach is to restrict the estimator to be **linear in the data** and find the linear estimator that is **unbiased and has minimum variance**. This estimator is termed the **best linear unbiased estimator (BLUE)**.

5.1 BLUE for a Scalar Parameter

The **BLUE** restricts the estimator to be linear class in the form of

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T \mathbf{x}$$

where $\mathbf{a} \in \mathbb{R}^{N \times 1}$ is to be determined to **minimize the variance**.

Notice that the **MVU is unbiased**. In order to meet the **unbiased criterion**, it must satisfy

$$E[\hat{\theta}] = \mathbf{a}^T E[\mathbf{x}] = \theta$$

and we need $E[\mathbf{x}] = \theta \mathbf{s}$ and $\mathbf{a}^T \mathbf{s} = 1$. To make life easier

$$\begin{aligned} \mathbf{x} &= E[\mathbf{x}] + \mathbf{x} - E[\mathbf{x}] \\ &= \theta \mathbf{s} + \mathbf{w} \end{aligned}$$

where $E[\mathbf{w}] = \mathbf{0}$ and $E[\mathbf{w}\mathbf{w}^T] = \mathbf{C}$. The variance of the linear estimator is

$$\begin{aligned} \text{var}(\hat{\theta}) &= E \left[\left(\sum_{n=0}^{N-1} a_n x[n] - E \left[\sum_{n=0}^{N-1} a_n x[n] \right] \right)^2 \right] \\ &= E \left[(\mathbf{a}^T \mathbf{x} - \mathbf{a}^T E[\mathbf{x}])^2 \right] \\ &= \mathbf{a}^T E[(\mathbf{x} - E[\mathbf{x}]) (\mathbf{x} - E[\mathbf{x}])^T] \mathbf{a} \\ &= \mathbf{a}^T \mathbf{C} \mathbf{a} \end{aligned}$$

To minimize the variance,

$$\min_{\mathbf{a}} \mathbf{a}^T \mathbf{C} \mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^T \mathbf{s} = 1.$$

Let $\mathbf{C}$ is positive definite and has a decomposition

$$\mathbf{C}^{-1} = \mathbf{D}^T \mathbf{D}.$$

All vector $\mathbf{a}$ can be written as $\mathbf{a} = \mathbf{D}^T \tilde{\mathbf{a}}$ and we can instead solve for $\tilde{\mathbf{a}}$. Since $\mathbf{C} = \mathbf{D}^{-1} \mathbf{D}^{-T}$, we have $\mathbf{a}^T \mathbf{C} \mathbf{a} = \tilde{\mathbf{a}}^T \mathbf{D} \mathbf{C} \mathbf{D}^T \tilde{\mathbf{a}} = \|\tilde{\mathbf{a}}\|^2$ and $\mathbf{a}^T \mathbf{s} = \tilde{\mathbf{a}}^T \mathbf{D} \mathbf{s}$.

$$\min_{\tilde{\mathbf{a}}} \|\tilde{\mathbf{a}}\|^2 \quad \text{s.t.} \quad \tilde{\mathbf{a}}^T \mathbf{D} \mathbf{s} = 1.$$

Now, any $\tilde{\mathbf{a}}$ such that $\tilde{\mathbf{a}}^T \mathbf{D} \mathbf{s} = 1$ can be written as

$$\tilde{\mathbf{a}} = \frac{\mathbf{D} \mathbf{s}}{\|\mathbf{D} \mathbf{s}\|^2} + \mathbf{v}, \quad (4)$$

where $\mathbf{v}^T \mathbf{D} \mathbf{s} = 0$. Determining $\mathbf{v}$ to minimize $\|\tilde{\mathbf{a}}\|^2 = \left\| \frac{\mathbf{D} \mathbf{s}}{\|\mathbf{D} \mathbf{s}\|^2} + \mathbf{v} \right\|^2$ is of interest.

$$\left\| \frac{\mathbf{D} \mathbf{s}}{\|\mathbf{D} \mathbf{s}\|^2} + \mathbf{v} \right\|^2 = \frac{1}{\|\mathbf{D} \mathbf{s}\|^2} + \|\mathbf{v}\|^2 + 2\mathbf{v}^T \mathbf{D} \mathbf{s} / \|\mathbf{D} \mathbf{s}\|^2 = \frac{1}{\|\mathbf{D} \mathbf{s}\|^2} + \|\mathbf{v}\|^2.$$

Thus, for all $\mathbf{v}$ subject to $\mathbf{v}^T \mathbf{D} \mathbf{s} = 0$,

$$\|\tilde{\mathbf{a}}\|^2 \geq \frac{1}{\|\mathbf{D} \mathbf{s}\|^2}$$

and this is achieved when $\mathbf{v} = \mathbf{0}$. Thus, the **optimal linear estimator (i.e., BLUE)** is

$$\begin{aligned} \tilde{\mathbf{a}}_{opt} &= \frac{\mathbf{D} \mathbf{s}}{\|\mathbf{D} \mathbf{s}\|^2} = \frac{\mathbf{D} \mathbf{s}}{\mathbf{s}^T \mathbf{D}^T \mathbf{D} \mathbf{s}} = \frac{\mathbf{D} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} \\ \mathbf{a}_{opt} &= \mathbf{D}^T \tilde{\mathbf{a}}_{opt} = \frac{\mathbf{D}^T \mathbf{D} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} = \frac{\mathbf{C}^{-1} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}. \end{aligned}$$

Notice that for the DC level estimation with $\mathbf{x} = \mathbf{1}A + \mathbf{w}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, the BLUE is

$$\mathbf{a}_{opt} = \frac{\mathbf{C}^{-1} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}} = \frac{\sigma^2 \mathbf{1}}{\sigma^2 \mathbf{1}^T \mathbf{1}} = \frac{1}{N} \mathbf{1}.$$

5.2 BLUE for a Vector Parameter

When $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_p]^T$, **linear estimation** has the structure

$$\hat{\boldsymbol{\theta}} = \mathbf{A} \mathbf{x}, \quad \mathbf{A} \in \mathbb{R}^{p \times N}$$

and $N \geq p$. To make it unbiased, we need

$$E \left[\hat{\boldsymbol{\theta}} \right] = \mathbf{A} E[\mathbf{x}] = \boldsymbol{\theta}.$$

For linear unbiased estimators to exist,

$$E[\mathbf{x}] = \mathbf{H} \boldsymbol{\theta}, \quad \mathbf{H} \in \mathbb{R}^{N \times p}$$

and

$$E \left[\hat{\boldsymbol{\theta}} \right] = \mathbf{A} \mathbf{H} \boldsymbol{\theta} = \boldsymbol{\theta} \Rightarrow \mathbf{A} \mathbf{H} = \mathbf{I}_p.$$

To make this possible,

$$\mathbf{x} = E[\mathbf{x}] + \mathbf{x} - E[\mathbf{x}] = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

and $E[\mathbf{w}] = \mathbf{0}$ and $E[\mathbf{w}\mathbf{w}^T] = \mathbf{C}$. The estimation error covariance is

$$\begin{aligned} \mathbf{C}_{\hat{\boldsymbol{\theta}}} &= E \left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)^T \right] \\ &= E \left[\mathbf{A} (\mathbf{x} - E[\mathbf{x}]) (\mathbf{x} - E[\mathbf{x}])^T \mathbf{A}^T \right] \\ &= \mathbf{A} \mathbf{C} \mathbf{A}^T \end{aligned}$$

For convenience, we again let $\mathbf{C}^{-1} = \mathbf{D}^T \mathbf{D}$ and $\mathbf{C} = \mathbf{D}^{-1} \mathbf{D}^{-T}$. Instead of solving for $\mathbf{A}$, we solve for $\tilde{\mathbf{A}}$ where

$$\mathbf{A} = \tilde{\mathbf{A}} \mathbf{D}.$$

This yields $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \tilde{\mathbf{A}} \mathbf{D} \mathbf{C} \mathbf{D}^T \tilde{\mathbf{A}}^T = \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T$ and the constraint $\mathbf{A} \mathbf{H} = \mathbf{I}_p$ becomes $\tilde{\mathbf{A}} (\mathbf{D} \mathbf{H}) = \mathbf{I}_p$. Thus, all $\tilde{\mathbf{A}}$ satisfying $\tilde{\mathbf{A}} (\mathbf{D} \mathbf{H}) = \mathbf{I}_p$ can be written as

$$\tilde{\mathbf{A}} = (\mathbf{H}^T \mathbf{D}^T \mathbf{D} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T + \mathbf{V}, \quad \mathbf{V} \in \mathbb{R}^{p \times N}$$

where $\mathbf{V}(\mathbf{D} \mathbf{H}) = \mathbf{0}$.

To minimize the variance,

$$\min_{\tilde{\mathbf{A}}} \text{tr} \left(\tilde{\mathbf{A}} \tilde{\mathbf{A}}^T \right) \quad \text{s.t.} \quad \mathbf{V}(\mathbf{D} \mathbf{H}) = \mathbf{0}.$$

Determining $\mathbf{V}$ to minimize $\text{tr} \left(\tilde{\mathbf{A}} \tilde{\mathbf{A}}^T \right)$ is of interest.

$$\begin{aligned} \text{tr} \left(\left((\mathbf{H}^T \mathbf{D}^T \mathbf{D} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T + \mathbf{V} \right) \left((\mathbf{H}^T \mathbf{D}^T \mathbf{D} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T + \mathbf{V} \right)^T \right) &= \text{tr} \left((\mathbf{H}^T \mathbf{D}^T \mathbf{D} \mathbf{H})^{-1} + \mathbf{V} \mathbf{V}^T \right) \\ &= \text{tr} \left((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} + \mathbf{V} \mathbf{V}^T \right) \\ &\geq \text{tr} \left((\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \right). \end{aligned}$$

Thus, for all $\mathbf{V}(\mathbf{D} \mathbf{H}) = \mathbf{0}$, the above inequality is achieved as the equality when $\mathbf{V} = \mathbf{0}$.

$$\begin{aligned} \tilde{\mathbf{A}}_{opt} &= (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T \\ \mathbf{A}_{opt} &= \tilde{\mathbf{A}}_{opt} \mathbf{D} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{D}^T \mathbf{D} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \end{aligned}$$

The estimator error covariance is

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{A} \mathbf{C} \mathbf{A}^T = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1}.$$

If the data are truly Gaussian, then the BLUE is also the MVU estimator.

Theorem 5 (***Gauss-Markov Theorem***) Suppose the observations are of the general linear form

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where $\mathbf{H} \in \mathbb{R}^{N \times p}$ is a known matrix, $\boldsymbol{\theta} \in \mathbb{R}^{p \times 1}$ is the parameter vector to be estimated, and $\mathbf{w} \in \mathbb{R}^{N \times 1}$ is the noise vector with *zero mean and covariance $\mathbf{C}$* (the PDF of $\mathbf{w}$ is unknown), then the BLUE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$$

and the minimum variance of $\hat{\theta}_i$ is

$$\text{var}(\hat{\theta}_i) = \left[(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \right]_{ii}.$$

The covariance matrix of $\hat{\boldsymbol{\theta}}$ is

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = (\mathbf{H}^T \mathbf{C} \mathbf{H})^{-1}.$$

6 Appendices

6.1 Proof of Theorem 1

First we show the forward direction. Assume the factorization is true, i.e.,

$$p(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta) h(\mathbf{x}).$$

Then, for a given observation $T(\mathbf{x}) = t$, we can write

$$\int_{\mathbf{x} \in \{\mathbf{x} | T(\mathbf{x}) = t\}} p(\mathbf{x}; \theta) d\mathbf{x} = p(T(\mathbf{x}) = t; \theta).$$

Thus, we can write

$$\begin{aligned} p(T(\mathbf{x}) = t; \theta) &= \int_{\mathbf{x} \in \{\mathbf{x} | T(\mathbf{x}) = t\}} g(T(\mathbf{x}); \theta) h(\mathbf{x}) d\mathbf{x} \\ &= g(T(\mathbf{x}) = t; \theta) \int_{\mathbf{x} \in \{\mathbf{x} | t = T(\mathbf{x})\}} h(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Furthermore,

$$\begin{aligned} p(\mathbf{x} | T(\mathbf{x}) = t; \theta) &= \frac{p(\mathbf{x}, T(\mathbf{x}) = t; \theta)}{p(T(\mathbf{x}) = t, \theta)} \\ &= \frac{g(T(\mathbf{x}); \theta) h(\mathbf{x}) \delta(T(\mathbf{x}) = t)}{g(T(\mathbf{x}) = t; \theta) \int_{\mathbf{x} \in \{\mathbf{x} | t = T(\mathbf{x})\}} h(\mathbf{x}) d\mathbf{x}} \\ &= \frac{h(\mathbf{x}) \delta(T(\mathbf{x}) = t)}{\int_{\mathbf{x} \in \{\mathbf{x} | t = T(\mathbf{x})\}} h(\mathbf{x}) d\mathbf{x}}, \end{aligned}$$

implying $p(\mathbf{x} | T(\mathbf{x}) = t; \theta)$ is independent of θ . Thus, $T(\mathbf{x})$ is sufficient.

The converse can be shown by assuming $T(\mathbf{x})$ is sufficient statistics and thus, $p(\mathbf{x} | T(\mathbf{x}) = t; \theta) = p(\mathbf{x} | T(\mathbf{x}) = t)$. Then, we can write

$$\begin{aligned} p(\mathbf{x}; \theta) \delta(T(\mathbf{x}) = t) &= p(\mathbf{x} | T(\mathbf{x}) = t; \theta) p(T(\mathbf{x}) = t; \theta) \\ &= p(\mathbf{x} | T(\mathbf{x}) = t) p(T(\mathbf{x}) = t; \theta) \\ &= h(\mathbf{x}) g(T(\mathbf{x}); \theta) \delta(T(\mathbf{x}) = t). \end{aligned}$$

$$\Rightarrow p(\mathbf{x}; \theta) = h(\mathbf{x}) g(T(\mathbf{x}); \theta)$$

6.2 Proof of Theorem 2

Given that $\check{\theta}$ satisfies $E[\check{\theta}] = \theta$ and $T(\mathbf{x})$ is sufficient statistics, we have

$$p(\mathbf{x} | T(\mathbf{x}); \theta) = p(\mathbf{x} | T(\mathbf{x})).$$

- 1) We show $\hat{\theta}$ is a solely a function of $T(\mathbf{x})$.

$$\begin{aligned}\hat{\theta} &= E[\check{\theta}|T(\mathbf{x})] \\ &= \int \check{\theta} p(\mathbf{x}|T(\mathbf{x}); \theta) d\mathbf{x} \\ &= \int \check{\theta} p(\mathbf{x}|T(\mathbf{x})) d\mathbf{x}\end{aligned}$$

This verifies that $\hat{\theta}$ is solely a function of $T(\mathbf{x})$.

- 2) Unbiasedness can be checked by

$$\begin{aligned}E_T[\hat{\theta}] &= \int E[\check{\theta}|T(\mathbf{x})] p(T(\mathbf{x}); \theta) dT \\ &= \int \int \check{\theta} p(\mathbf{x}|T(\mathbf{x}); \theta) p(T(\mathbf{x}); \theta) d\mathbf{x} dT \\ &= \int \check{\theta} \int p(\mathbf{x}, T(\mathbf{x}); \theta) dT d\mathbf{x} \\ &= \int \check{\theta} p(\mathbf{x}; \theta) d\mathbf{x} \\ &= \theta.\end{aligned}$$

- 3) Need to show $\text{var}(\check{\theta}) \geq \text{var}(\hat{\theta})$

$$\begin{aligned}\text{var}(\check{\theta}) &= E[(\check{\theta} - \theta)^2] \\ &= E[(\check{\theta} - \hat{\theta} + \hat{\theta} - \theta)^2] \\ &= E[(\check{\theta} - \hat{\theta})^2] + E[(\hat{\theta} - \theta)^2] - 2E[(\check{\theta} - \hat{\theta})(\hat{\theta} - \theta)].\end{aligned}$$

The cross terms are

$$\begin{aligned}E[(\check{\theta} - \hat{\theta})(\hat{\theta} - \theta)] &= E[(\check{\theta} - \hat{\theta})\hat{\theta}] - E[(\check{\theta} - \hat{\theta})\theta] \\ &= E_T[E_{\mathbf{x}|T}[(\check{\theta} - \hat{\theta})\hat{\theta}|T(\mathbf{x})]] \\ &= E_T[(E_{\mathbf{x}|T}[\check{\theta}|T(\mathbf{x})] - \hat{\theta})\hat{\theta}] = 0\end{aligned}$$

because $E_{\mathbf{x}|T}[\check{\theta}|T(\mathbf{x})] = \hat{\theta}$. Thus,

$$\text{var}(\check{\theta}) = E[(\check{\theta} - \hat{\theta})^2] + \text{var}(\hat{\theta}) \geq \text{var}(\hat{\theta})$$

- 4) The estimator $\hat{\theta}$ is a function of $T(\mathbf{x})$ such that $\hat{\theta} = g(T(\mathbf{x})) = E[\check{\theta}|T(\mathbf{x})]$. This is also unbiased $E[\hat{\theta}] = \theta$. If $T(\mathbf{x})$ is complete, then we know only one such function exist such

that $\text{var}(g(T(\mathbf{x}))) \leq \text{var}(\check{\theta})$. Notice that this holds for any arbitrary unbiased $\check{\theta}$. Thus, $g(T(\mathbf{x})) = E[\check{\theta}|T(\mathbf{x})]$ is the MVU estimator.