Estimator Performance

Aleksandar Dogandžić

2017-01-19

Contents

Reading: §1–2, in the textbook and (Hero 2015, §4.4.1).

Continue with the DC-level-in-Gaussian noise estimation example from handout `intro`. Recall that the noise sequence $(W[n])$ is independent, identically distributed (i.i.d.).

Consider the following two estimators:[1]

$$\hat{a}_1(\boldsymbol{X}) = \frac{1}{N}\sum_{n=0}^{N-1} X[n]$$
$$\hat{a}_2(X[0]) = X[0]$$

where $\boldsymbol{X} = (X[n])_{n=0}^{N-1}$.

Which of the two estimators is better?

For a given realization of the measurements $\boldsymbol{X} = \boldsymbol{x} = (x[n])_{n=0}^{N-1}$, it is possible that either $\hat{a}_1(\boldsymbol{x})$ or $\hat{a}_2(x[0])$ is closer to a. Hence, we need statistical analysis to answer this question.

Substitute the measurement model to perform statistical analysis:

$$\hat{a}_1(\boldsymbol{X}) = \frac{1}{N}\sum_{n=0}^{N-1}(\underbrace{a + W[n]}_{X[n]}) \tag{1a}$$

$$\hat{a}_2(X[0]) = \underbrace{a + W[0]}_{X[0]}\ . \tag{1b}$$

Take expectations:

$$\mathrm{E}_{\boldsymbol{X}\,|\,A}\big[\hat{a}_1(\boldsymbol{X})\,|\,a\big] = \frac{1}{N}\sum_{n=0}^{N-1}\{a + \underbrace{\mathrm{E}_W\big(W[n]\big)}_{0}\}$$
$$= a \tag{2a}$$

$$\mathrm{E}_{X\,|\,A}\big[\hat{a}_2(X[0])\,|\,a\big] = a + \underbrace{\mathrm{E}_W\big(W[0]\big)}_{0}$$
$$= a. \tag{2b}$$

[1] Here, $\hat{a}_1(\boldsymbol{X})$ is the maximum-likelihood (ML) estimate of the DC level a, i.e., it maximizes the likelihood function of a. Interestingly, the ML estimate of a is the same regardless of whether σ^2 is known or not. It is also intuitively appealing: a is the *average level* of $X[n]$ because $W[n]$ has zero mean.

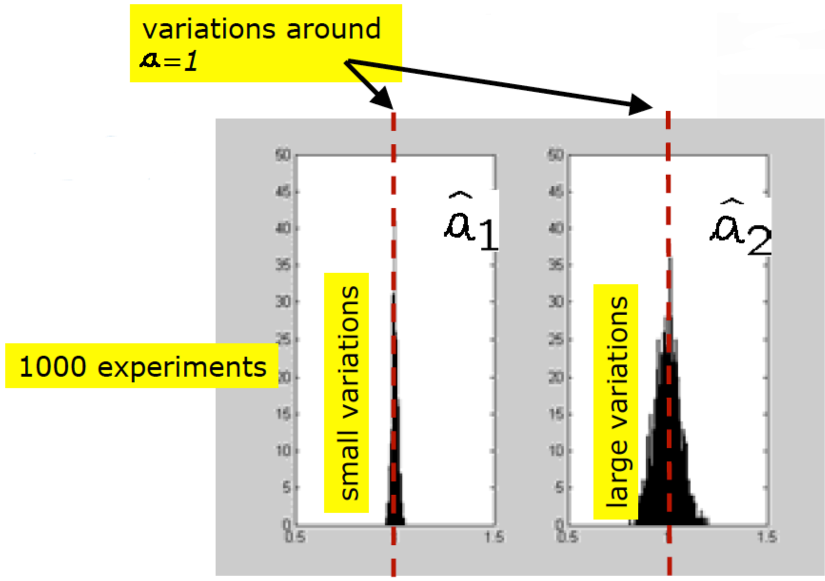

Figure 1: Histograms of $\hat{a}_1(\boldsymbol{X})$ and $\hat{a}_2(\boldsymbol{X})$ obtained using 1000 realizations of $\boldsymbol{X}$ and the DC level $a = 1$.

On average, both estimators are centered around the correct value, i.e., they are *unbiased*.

✻ SIMPLIFIED notation. In the *classical scenario* that we describe here, the DC level a is a constant, which is why $\mathrm{E}_{\boldsymbol{X}|A}(\cdot \,|\, a)$ is equivalent to $\mathrm{E}_{\boldsymbol{X}}(\cdot)$. When it is clear from context that we are dealing with the classical scenario and that a is the value of the parameter, we may simplify the notation and use

$$\mathrm{E}_{\boldsymbol{X}}(\cdot). \tag{3a}$$

✻ MORE simplified notation:

$$\hat{a}_i = \hat{a}_i(\boldsymbol{X}). \tag{3b}$$

But, $\hat{a}_1$ is *better* than $\hat{a}_2$ because its probability density function (pdf) is more concentrated around the true value, see Fig. 1. On average, $\hat{a}_1$ is closer to the true a.

Proof: The mean values are

$$\mathrm{E}_{\boldsymbol{X}}(\hat{a}_1) = \mathrm{E}_{\boldsymbol{X}}(\hat{a}_2) = a$$

and the variances are

$$\begin{aligned}
\mathrm{var}_{\boldsymbol{X}}(\hat{a}_1) &= \mathrm{E}_{\boldsymbol{X}}\{[\hat{a}_1 - \mathrm{E}_{\boldsymbol{X}}(\hat{a}_1)]^2\} \\
&= \mathrm{E}_{\boldsymbol{W}}\Big\{\Big(\frac{1}{N}\sum_{n=0}^{N-1} W[n]\Big)^2\Big\} \\
&= \frac{1}{N^2}\sum_{n=0}^{N-1} \mathrm{var}_W(W[n]) \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

$\mathrm{E}_{\boldsymbol{X}}(\hat{a}_1) = a$

$W[n]$ i.i.d.

and

$$\begin{aligned}
\mathrm{var}_X(\hat{a}_2) &= \mathrm{var}_X(X[0]) \\
&= \mathrm{var}_W(W[0]) \\
&= \sigma^2. \qquad \square
\end{aligned}$$

Which estimator is the best? To answer this question, we need to select a criterion that measures estimation quality.

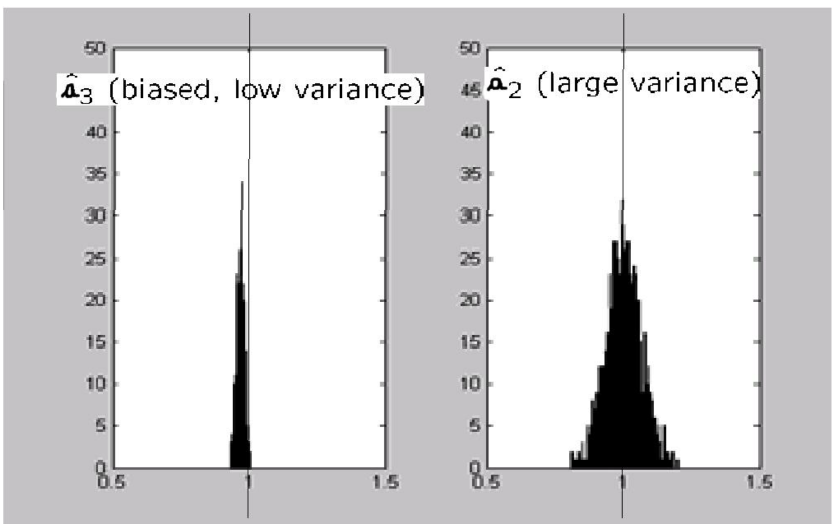

Figure 2: Which estimator is better?

Definition 1. Bias and mean-square error (MSE) of an estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ under the classical setting:

$$\mathrm{bias}(\hat{\theta}) = b(\theta) = \mathrm{E}_{\boldsymbol{X}}(\hat{\theta}) - \theta \tag{3a}$$

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_{\boldsymbol{X}}[(\hat{\theta} - \theta)^2] \\
&= \mathrm{var}_{\boldsymbol{X}}(\hat{\theta}) + b^2(\theta).
\end{aligned} \tag{3b}$$

We wish to minimize the above MSE, which may lead to a minimum mean-square error (MMSE) estimator. We now show the second equality in (3b). Define

$$\mu(\theta) \triangleq \mathrm{E}_{\boldsymbol{X}}(\hat{\theta}).$$

The MSE can be written as

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_{\boldsymbol{X}}[(\hat{\theta}-\theta)^2] \\
&= \mathrm{E}_{\boldsymbol{X}}\left(\left[\hat{\theta}-\mathrm{E}_{\boldsymbol{X}}(\hat{\theta})+\mathrm{E}_{\boldsymbol{X}}(\hat{\theta})-\theta\right]^2\right) \\
&= \mathrm{E}_{\boldsymbol{X}}\left\{[\hat{\theta}-\mu(\theta)+\mu(\theta)-\theta]^2\right\} \\
&= \underbrace{\mathrm{E}_{\boldsymbol{X}}\left\{[\hat{\theta}-\mu(\theta)]^2\right\}}_{\mathrm{var}_{\boldsymbol{X}}(\hat{\theta})} + \underbrace{[\mu(\theta)-\theta]^2}_{b^2(\theta)} \\
&\quad \underbrace{+2\,\mathrm{E}_{\boldsymbol{X}}\left\{[\hat{\theta}-\mu(\theta)]\overbrace{[\mu(\theta)-\theta]}^{\text{const}}\right\}}_{0}\,. \qquad (4)
\end{aligned}$$

In the classical scenario, minimizing the MSE would be a reasonable estimator design criterion. However, $\mathrm{MSE}(\hat{\theta}(\boldsymbol{X}))$ is usually a function of θ: minimizing it over some family of estimators will often produce an optimal "estimator" $\hat{\theta}(\boldsymbol{X})$ that depends on θ.

Example: DC level in AWGN

see §2.4 in the textbook

CONSIDER

$$X[n] = a + W[n]$$
$$(W[n])_{n=0}^{N-1} \quad \text{i.i.d. } \mathcal{N}(0,\sigma^2).$$

$W[n]$ is additive white Gaussian noise (AWGN).

Consider the following family of estimators of a:

$$\tilde{a} = c\overline{X} \qquad (5)$$

where

$$\overline{X} = \frac{1}{N}\sum_{n=0}^{N-1} X[n].$$

sample mean

Here

$$\mathrm{E}_{\boldsymbol{X}}(\tilde{a}) = ca, \quad \mathrm{var}_{\boldsymbol{X}}(\tilde{a}) = \frac{c^2\sigma^2}{N}.$$

Find the best c that minimizes the MSE for the estimator family (5) of the DC parameter a. In other words, can we improve upon the sample mean?

$$\mathrm{MSE}\{\tilde{a}\} = \mathrm{MSE}(c) = \frac{c^2\sigma^2}{N} + \overbrace{(ca-a)^2}^{\text{depends on } a}$$
$$\frac{\mathrm{d}\,\mathrm{MSE}(c)}{\mathrm{d}c} = \frac{2c\sigma^2}{N} + 2(ca-a)a = 0$$

and

$$c_{\mathrm{opt}} = \frac{a^2}{a^2+\sigma^2/N}$$

depends on the unknown parameter a. Hence not useful. Observe the shrinkage form of $\tilde{a}$:

$$c_{\text{opt}} \in [0, 1]$$

and, consequently, $\tilde{a}$ has smaller magnitude than $\overline{X}$.

Admissible and MVU Estimators

AN estimator $\hat{\theta}$ is *inadmissible* (for MSE as a performance measure) if there exists another estimator $\tilde{\theta}$ that dominates it, i.e., if $\text{MSE}(\tilde{\theta}) \leq \text{MSE}(\hat{\theta})$ for all θ in the parameter space sp_Θ ($\forall \theta \in \text{sp}_\Theta$), with strict inequality for certain θ. An estimator is *admissible* otherwise.

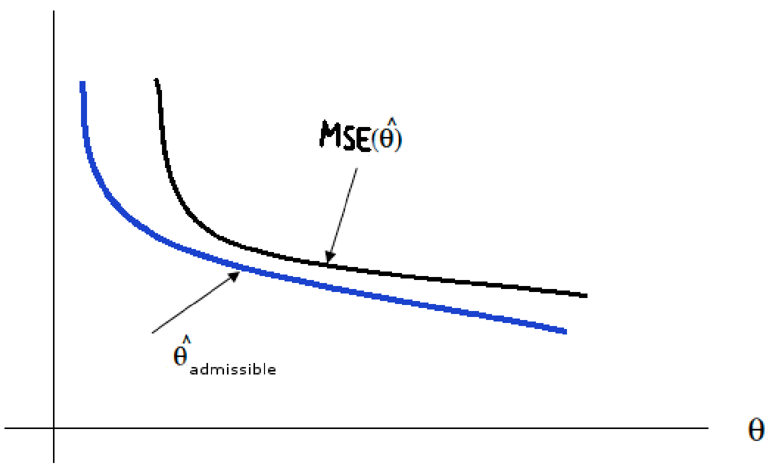

Figure 3: An admissible estimator $\hat{\theta}_{\text{admissible}}$ has lower (or equal) MSE than any other estimator $\hat{\theta}$ for all $\theta \in \text{sp}_\Theta$.

We may focus on the family of estimators that have zero bias and then minimize estimator variance[2] for all values of θ: minimum-variance unbiased (MVU) estimator $\hat{\theta}_{\text{MVU}} = \hat{\theta}_{\text{MVU}}(\boldsymbol{x})$.

MVU estimator $\hat{\theta}_{\text{MVU}}$ *does not* always exist, since $\hat{\theta}_{\text{MVU}}$ must have the smallest variance for all values of θ,[3] see Fig. 5.

[2] Since the bias is zero, the estimator variance is equal to its MSE for this family.

[3] To emphasize that the MVU estimator must have the smallest variance for all values of θ, (Hero 2015) refers to it as uniformly minimum-variance unbiased (UMVU).

✻ COMMENTS:

- Even if it exists for a particular problem, MVU estimator is not optimal in terms of minimizing the MSE and we may be able to do better. Usually, the MVU estimator is not admissible in terms of MSE, but in special cases it is.

- By relaxing the unbiasedness condition, it is possible to outperform the MVU estimators in MSE, as shown in the following example.

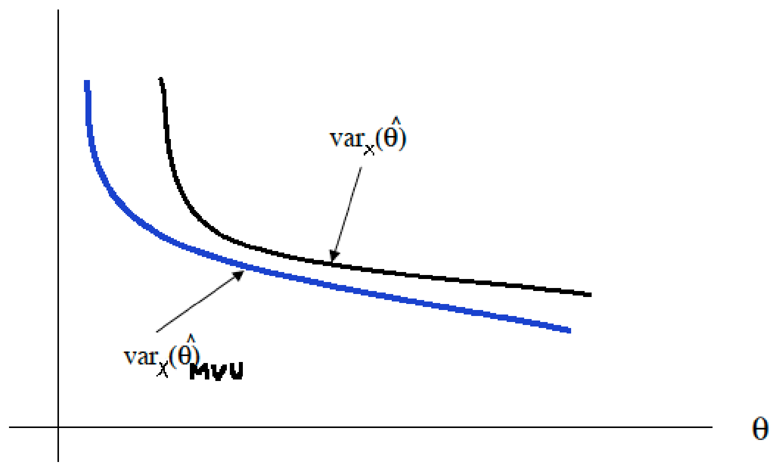

Figure 4: An MVU estimator $\hat{\theta}_{\text{MVU}}$ is an unbiased estimator that has lower variance than any other unbiased estimator $\hat{\theta}$ for all θ.

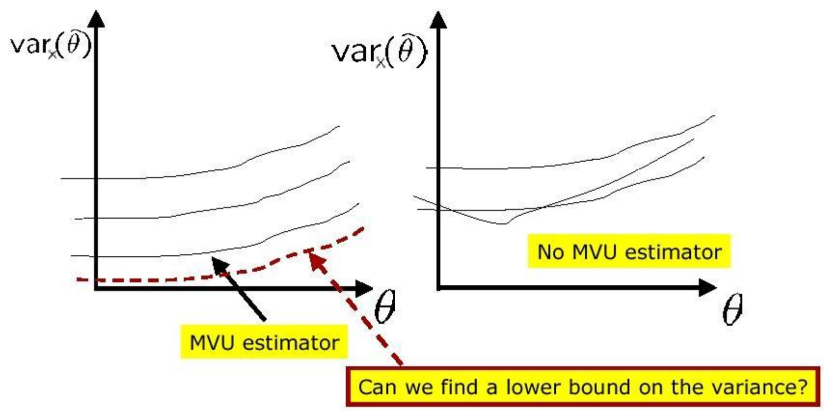

Figure 5: Variances of unbiased parameter estimators as functions of the parameter value.

✻ EXAMPLE: Variance estimation. Consider now estimating the variance σ^2 of i.i.d. zero-mean Gaussian observations, using the following family of estimators of σ^2:

This example is based on (Stoica and Moses 1990).

$$\hat{\sigma}^2 = \hat{\sigma}^2(\boldsymbol{X}) = c\,\frac{1}{N}\sum_{n=0}^{N-1} X^2[n] \tag{6}$$

where $c > 0$ is a constant that we need to select. If we choose $c = 1$, $\hat{\sigma}^2\big|_{c=1}$ is an unbiased estimator with

$$\hat{\sigma}^2\big|_{c=1} = \hat{\sigma}^2_{\text{MVU}} = \frac{1}{N}\sum_{n=0}^{N-1} X^2[n]. \tag{7}$$

Now, for the estimator family (6),

$$\mathrm{E}_{\boldsymbol{X}}(\hat{\sigma}^2) = c\sigma^2$$

and

$$\begin{aligned}
\mathrm{MSE}\{\hat{\sigma}^2\} &= \mathrm{E}_{\boldsymbol{X}}\left[(\hat{\sigma}^2-\sigma^2)^2\right] && (8)\\
&= \mathrm{E}_{\boldsymbol{X}}\left[(\hat{\sigma}^2)^2\right] + (\sigma^2)^2 - 2\sigma^2\,\mathrm{E}_{\boldsymbol{X}}(\hat{\sigma}^2)\\
&= \mathrm{E}_{\boldsymbol{X}}\left[(\hat{\sigma}^2)^2\right] + (\sigma^2)^2(1-2c)\\
&= \frac{c^2}{N^2}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}\mathrm{E}_{\boldsymbol{X}}(X^2[n_1]X^2[n_2]) + (\sigma^2)^2(1-2c)\\
&= \frac{c^2}{N^2}\Big\{(N^2-N)(\sigma^2)^2 + N\,\underbrace{\mathrm{E}_{\boldsymbol{X}}(X^4[n])}_{3(\sigma^2)^2}\Big\} + (\sigma^2)^2(1-2c)\\
&= (\sigma^2)^2\left[c^2\left(1+\frac{2}{N}\right)+(1-2c)\right] && (9)
\end{aligned}$$

where we have used the following facts:

- For $n_1 \neq n_2$,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{X}}\left(X^2[n_1]X^2[n_2]\right) &= \mathrm{E}_{X}\left(X^2[n_1]\right)\mathrm{E}_{X}(X^2[n_2])\\
&= \sigma^2\sigma^2\\
&= (\sigma^2)^2. && (10)
\end{aligned}$$

- For $n_1 = n_2$,

$$\begin{aligned}
\mathrm{E}_{X}\left(X^2[n_1]X^2[n_1]\right) &= \mathrm{E}_{X}\left(X^4[n_1]\right)\\
&= 3(\sigma^2)^2 && (11)
\end{aligned}$$

  which is the fourth-order moment of a Gaussian distribution having zero mean and variance σ^2.

Here, (9) is minimized for

$$c_{\mathrm{OPT}} = \frac{N}{N+2}$$

yielding the estimator

$$\hat{\sigma}^2_\star = c_{\mathrm{OPT}}\frac{1}{N}\sum_{n=0}^{N-1}X^2[n]$$

whose MSE is minimum for the family of estimators in (6):

$$\mathrm{MSE}_{\mathrm{MIN}} = \frac{2(\sigma^2)^2}{N+2}.$$

✻ COMMENTS:

- $\hat{\sigma}^2_\star$ is *biased* and has *smaller MSE* than the MVU estimator in (7):

$$\mathrm{MSE}_{\mathrm{MIN}} < \mathrm{MSE}\{\hat{\sigma}^2\}\big|_{c=1} = \frac{2\sigma^4}{N}.$$

- Unlike in the previous example, we are able to construct a realizable estimator in this case.
- For large N, $\hat{\sigma}_\star^2$ and $\hat{\sigma}_{\text{MVU}}^2$ are approximately the same because $\frac{N}{N+2} \nearrow 1$ as $N \nearrow \infty$. This also implies that $\hat{\sigma}_\star^2$ is *asymptotically unbiased*.

References

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI.

Stoica, Petre and Randolph L. Moses (1990). "On biased estimators and the unbiased Cramér-Rao lower bound". In: *Signal Process.* 21.4, pp. 349–350.