ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak , **scribe:** Inseok Heo

Lecture 17: Minimum Variance Unbiased (MVUB) Estimators

Ultimately, we would like to be able to argue that a given estimator is(or is not) optimal in some sense. Usually this is very difficult, but in certain cases it is possible to make precise statements about optimality. MVUB estimators are one class where this is sometimes the case. First let's quickly review some key estimation-theoretic concepts.

1 Review of key estimation concepts

Observation model

$$X \sim p(x|\theta), \quad x \in \mathcal{X}, \quad \theta \in \Theta$$

Loss

$$\ell : \Theta \times \Theta \to \mathbb{R}_+$$

Example 1

$$\begin{aligned} \ell_2 \quad &: \quad \ell(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2 \\ \ell_1 \quad &: \quad \ell(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_1 \\ \text{log-likelihood} \quad &: \quad \ell(\theta_1, \theta_2) = -\log p(x|\theta_2), \quad \text{where } x \sim p(x|\theta_1) \end{aligned}$$

Risk : expected loss of estimator $\hat{\theta}(x)$

$$R(\theta^*, \hat{\theta}) = \mathbb{E}[\ell(\theta^*, \hat{\theta})]$$

MSE : the ℓ_2 risk is usually called the mean square error :

$$MSE(\hat{\theta}) = \mathbb{E}[\|\theta^* - \hat{\theta}\|_2^2]$$

Recall that the *MSE* can be decomposed into the bias and variance

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[\|\theta^* - \hat{\theta}\|_2^2] \\ &= \mathbb{E}[\|\theta^* - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta}\|_2^2] \\ &= \|\theta^* - \mathbb{E}\hat{\theta}\|_2^2 + 2\mathbb{E}[(\theta^* - \mathbb{E}\hat{\theta})^T(\mathbb{E}\hat{\theta} - \hat{\theta})] + \mathbb{E}[\|\mathbb{E}\hat{\theta} - \hat{\theta}\|_2^2] \\ &= \|\theta^* - \mathbb{E}\hat{\theta}\|_2^2 + \mathbb{E}[\|\hat{\theta} - \mathbb{E}\hat{\theta}\|_2^2] \\ &= \text{bias}^2(\hat{\theta}) + \text{var}(\hat{\theta}) \end{aligned}$$

It is usually impossible to design $\hat{\theta}$ to minimize the *MSE* because the bias depends on θ^*, which is of course unknown. But suppose we restrict our attention to *unbiased* estimators; *i.e.*, $\hat{\theta}$ satisfying $\mathbb{E}\hat{\theta} = \theta^*$. Then

$$MSE(\hat{\theta}) \;=\; \text{var}(\hat{\theta})$$

and $\text{var}(\hat{\theta})$ does not depend on θ^*.

So a realizable approach is to optimize the *MSE* with respect to the class of unbiased estimators. The Minimum Variance UnBiased (MVUB) estimator is defined as

$$\tilde{\theta} \;=\; \underset{\hat{\theta} \,:\, \mathbb{E}[\hat{\theta}] = \theta^*}{\arg\min} \; \mathbb{E}[\|\hat{\theta} - \mathbb{E}\hat{\theta}\|_2^2]$$

Example 2 $X_1, X_2, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta^*, 1)$

$$\begin{aligned} \hat{\theta} &= \frac{1}{n}\sum x_i \;\Rightarrow\; \mathbb{E}[\hat{\theta}] \;=\; \theta^* \\ MSE(\hat{\theta}) &= \mathbb{E}\left[\left|\frac{1}{n}\sum x_i - \mathbb{E}\hat{\theta}\right|^2\right] \\ &= \mathrm{var}\left(\frac{1}{n}\sum x_i\right) \\ &= \frac{1}{n^2}\sum \mathrm{var}(x_i) \;=\; \frac{1}{n} \end{aligned}$$

Is this the *MVUB* estimator?

2 Finding the MVUB estimator

Finding the *MVUB* estimator can be difficult, but sometimes it is easy to verify that a particular estimator is *MVUB*.

Theorem 1 (*Cramér-Rao Lower Bound* (*CRLB*))
Let x denote an n-dimensional random vector with density $p(x|\theta^*)$, $\theta^* \in \mathbb{R}^k$
Assume that the first and second derivatives of $\log p(x|\theta)$ *exist.*
Let $\hat{\theta} = \hat{\theta}(x)$ be an unbiased estimator of θ^*. Then the error covariance *satisfies the matrix inequality*

$$\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})^T] \;\geq\; I^{-1}(\theta^*)$$

where $I(\theta^*)$ is the Fisher-Information matrix *with i,jth element*

$$I_{ij}(\theta^*) \;=\; -\mathbb{E}\left[\left.\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i \partial\theta_j}\right|_{\theta=\theta^*}\right]$$

Remark : The meaning of the inequality

$$C \;:=\; \mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})^T] \;\geq\; I^{-1}(\theta^*)$$

is that the eigenvalues of the symmetric matrix

$$C - I^{-1}(\theta^*)$$

are non-negative. As a consequence

$$\begin{aligned} \mathrm{var}(\hat{\theta}) &= \mathrm{tr}(\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\hat{\theta} - \mathbb{E}\hat{\theta})^T]) \\ &= \mathrm{tr}(C) \;\geq\; \mathrm{tr}(I^{-1}(\theta^*)) \end{aligned}$$

Proof : We will prove the scalar case ($\theta \in \mathbb{R}$). The general case follows in a similar fashion. The Fisher-Information is scalar in this case :

$$I(\theta^*) \;=\; -\mathbb{E}\left[\left.\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\right|_{\theta=\theta^*}\right]$$

Before proceeding we will show first that

$$I(\theta^*) \;=\; \mathbb{E}\left[\left.\left(\frac{\partial \log p(x|\theta)}{\partial\theta}\right)^2\right|_{\theta=\theta^*}\right]$$

To this end, first observe that

$$\begin{aligned}\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{\partial \log p(x|\theta)}{\partial\theta}\right) = \frac{\partial}{\partial\theta}\left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta^2}\right)\\ &= -\frac{1}{p^2(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta}\frac{\partial p(x|\theta)}{\partial\theta} + \frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial\theta^2}\\ &= -\left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta}\right)^2 + \frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial\theta^2}\end{aligned}$$

Consider the expectation of the second term :

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial\theta^2}\right] &= \int \frac{1}{p(x|\theta)}\frac{\partial^2 p(x|\theta)}{\partial\theta^2}p(x|\theta)dx\\ &= \int \frac{\partial^2 p(x|\theta)}{\partial\theta^2}dx\\ &= \frac{\partial^2}{\partial\theta^2}\int p(x|\theta)dx \;= \frac{\partial^2}{\partial\theta^2}(1)\\ &= 0\end{aligned}$$

Thus,

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2}\bigg|_{\theta=\theta^*}\right] &= \mathbb{E}\left[\left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta}\right)^2\bigg|_{\theta=\theta^*}\right]\\ &= \mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial\theta}\right)^2\bigg|_{\theta=\theta^*}\right]\end{aligned}$$

The gradient of the log-likelihood is called the *score function*. Let's denote it

$$S(\theta, x) \;:=\; \frac{\partial}{\partial\theta}\log p(x|\theta)$$

Observe that the MLE satisfies $S(\hat{\theta}, x) = 0$. Also note that

$$\mathbb{E}[S(\theta^*, x)] \;=\; \int \frac{\partial}{\partial\theta}\log p(x|\theta)|_{\theta=\theta^*} p(x|\theta^*)dx \;=\; \int \frac{\partial}{\partial\theta}p(x|\theta)dx \;=\; 0$$

and therefore the Fisher-Information is the variance of the *score function*

$$I(\theta^*) \;=\; \mathbb{E}\left[\left(\frac{\partial \log p(x|\theta)}{\partial\theta}\right)^2\bigg|_{\theta=\theta^*}\right]$$

So we see that the Fisher-Information measures the variability of the *score function* at $\theta = \theta^*$. We will also show that

$$\mathbb{E}[S(\theta^*, x)(\hat{\theta} - \theta^*)] = 1$$

To verify this, note that

$$\mathbb{E}[\hat{\theta} - \theta^*] = \int (\hat{\theta} - \theta^*)p(x|\theta^*)dx = 0$$

since $\hat{\theta}$ is unbiased. Take the derivative

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int(\hat{\theta}-\theta)p(x|\theta)dx\Big|_{\theta=\theta^*} \\
&= \left(-\int p(x|\theta)dx + \int(\hat{\theta}-\theta^*)\frac{\partial p(x|\theta)}{\partial\theta}dx\right)\Big|_{\theta=\theta^*} \\
&= -1+\int(\hat{\theta}-\theta^*)p(x|\theta)\frac{\partial\log p(x|\theta)}{\partial\theta}dx\Big|_{\theta=\theta*} \\
&= -1+\mathbb{E}[S(\theta^*,x)(\hat{\theta}-\theta^*)] \\
&\Rightarrow \mathbb{E}[S(\theta^*,x)(\hat{\theta}-\theta^*)]=1
\end{aligned}$$

Now we apply the *Cauchy-Schwarz inequality.*
(*i.e.,* $\int f(x)g(x)dx \le \sqrt{\int f^2(x)dx\int g^2(x)dx}$)

$$\begin{aligned}
1 &= \mathbb{E}[S(\theta^*,x)(\hat{\theta}-\theta^*)] \\
&\le \sqrt{\mathbb{E}[S^2(\theta^*,x)]}\sqrt{\mathbb{E}[(\hat{\theta}-\theta^*)^2]} \\
&= \sqrt{\mathrm{var}(S(\theta^*,x))}\sqrt{\mathrm{var}(\hat{\theta})} \\
\Rightarrow \mathrm{var}(\hat{\theta}) &\ge \frac{1}{\mathrm{var}(S(\theta^*,x))} = I^{-1}(\theta^*)
\end{aligned}$$

Example 3 $\quad X_1, X_2, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta^*,\,1)$

$$\begin{aligned}
\log p(x|\theta) &= \sum_{i=1}^n \log p(x_i|\theta) \\
\frac{\partial}{\partial\theta}\log p(x|\theta) &= \sum_{i=1}^n \frac{\partial}{\partial\theta}\log p(x_i|\theta) \\
&= \sum_{i=1}^n \frac{\partial}{\partial\theta}\log\left(\frac{1}{\sqrt{2\pi}}\; e^{\frac{-(x_i-\theta)^2}{2}}\right) \\
&= \sum_{i=1}^n (x_i-\theta)
\end{aligned}$$

$$\begin{aligned}
I(\theta^*) &= \mathbb{E}\left[\left(\frac{\partial\log p(x|\theta)}{\partial\theta}\right)^2\Big|_{\theta=\theta^*}\right] = \sum_{i=1}^n \mathbb{E}[(x_i-\theta^*)^2] = n \\
&\Rightarrow \quad \textit{MVUB} \text{ estimator variance} \ge \frac{1}{n}
\end{aligned}$$

But recall the unbiased estimator

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$$

$$\mathrm{var}(\widehat{\theta^2}) = \frac{1}{n} \quad \Rightarrow \quad \hat{\theta} \;\; \textit{is the} \text{ MVUB estimator!}$$

3 Efficiency

An unbiased estimator that achieved the *CRLB* is said to be *efficient*. Efficient estimators are *MVUB*, but not all *MVUB* estimators are necessarily efficient.

An estimator $\hat{\theta}_n$ is said to be *asymptotically efficient* if it achieves the *CRLB*, as $n \to \infty$.

Recall that under mild regularity conditions, the *MLE* has an asymptotic distribution

$$\hat{\theta}_n \overset{asymp}{\sim} \mathcal{N}(\theta^*, \frac{1}{n} I(\theta^*))$$

and so $\hat{\theta}_n$ is asymptotically unbiased and

$$\text{var}(\hat{\theta}_n) = \frac{1}{n} I^{-1}(\theta^*)$$

so it is *asymptotically efficient*.