ESE 524: Detection and Estimation Theory

Recitation 1

Washington University in St. Louis

Topics for Today

- Estimator performance
- Sufficiency

Useful Formulas

- Mean square error (MSE):

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + \mathrm{b}(\hat{\theta})^2$$

- Sufficient statistics:

$$p(\boldsymbol{x}|T(\boldsymbol{x}) = t; \theta) = p(\boldsymbol{x}|T(\boldsymbol{x}) = t)$$

- Factorization theorem:

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

Estimator Performance

- Estimator performance: We may have multiple estimators for the same problem. We are interested in finding the estimator that is optimal for a given criterion. The criteria will depend on the problem at hand.
- e.g., we may choose the estimator that minimizes the mean square error (MSE). Remember that $\mathrm{MSE}(\hat{\theta}) = \mathrm{var}(\hat{\theta}) + \mathrm{b}(\hat{\theta})^2$. However, this estimator usually depends on the real (unknown) parameters.
- We may choose the minimum variance unbiased (MVU) estimator. This estimator has the minimum variance of all the unbiased estimators. However, this estimator may not always exist.

Existence of the Minimum Variance Unbiased Estimator

- Assume that we have two independent observations $x[0]$ and $x[1]$ with p.d.f.

$$x[0] \sim \mathcal{N}(\theta, 1)$$

$$x[1] \sim \begin{cases} \mathcal{N}(\theta, 1), & if \theta \geq 0 \\ \mathcal{N}(\theta, 2), & if \theta < 0 \end{cases}$$

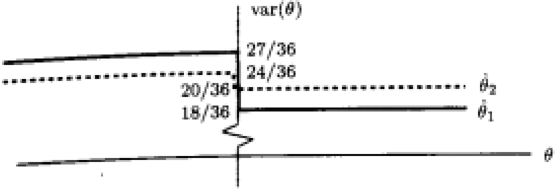
- The two estimators
  $\hat{\theta}_1 = \frac{1}{2}(x[0] + x[1])$
  $\hat{\theta}_2 = \frac{2}{3}x[0] + \frac{1}{2}x[1]$
  are unbiased. To see that:
  $\mathrm{E}(\hat{\theta}_1) = \frac{1}{2}\mathrm{E}(x[0] + x[1]) = \frac{1}{2}(\theta + \theta) = \theta$
  $\mathrm{E}(\hat{\theta}_2) = \frac{2}{3}\mathrm{E}x[0] + \frac{1}{2}\mathrm{E}x[1] = \theta$

Existence of the Minimum Variance Unbiased Estimator Cont.

Now, we calculate the variances and obtain:

$$\mathrm{var}(\hat{\theta}_1) = \frac{1}{4}(\mathrm{var}(x[0]) + \mathrm{var}(x[1])) = \begin{cases} \frac{18}{36}, & \text{if} \quad \theta \geq 0 \\ \frac{27}{36}, & \text{if} \quad \theta < 0 \end{cases}$$

$$\mathrm{var}(\hat{\theta}_2) = \frac{4}{9}\mathrm{var}(x[0]) + \frac{1}{9}\mathrm{var}(x[1]) = \begin{cases} \frac{20}{36}, & \text{if} \quad \theta \geq 0 \\ \frac{24}{36}, & \text{if} \quad \theta < 0 \end{cases}$$



Existence of the Minimum Variance Unbiased Estimator Cont.

- Between these two estimators no MVU estimator exists. In the next lecture we will learn about the Cramér-Rao bound and we will be able to show that $\hat{\theta}_1$ achieves the minimum possible variance of an unbiased estimator for $\theta \geq 0$ and $\hat{\theta}_2$ achieves the minimum possible variance of an unbiased estimator for $\theta < 0$.

Sufficient Statistics

- Let $\boldsymbol{x} = [x[0], x[1], x[2], ..., x[N-1]]^T$ be a set of samples, θ be an unknown parameter, and $p(\boldsymbol{x};\theta)$ be the corresponding probability density function.
- Statistic: Function $T(\boldsymbol{x})$ of the samples, intended to capture important information from the sample data.
- Sufficient Statistics: T is sufficient if $p(\boldsymbol{x}|T,\theta)$ is not a function of theta.
- One way to check sufficiency is using the definition of conditional densities:

$$p(\boldsymbol{x}|T,\theta) = \frac{p(\boldsymbol{x},T|\theta)}{p(T|\theta)}$$

- Instead of the factorization theorem, we will use this definition with Example 5.1 from Kay.

Normal pdf - Hard Version

- Let $\boldsymbol{x}$ consist of i.i.d. samples from a normal distribution with mean θ and known variance σ^2.
- Let $T(\boldsymbol{x}) = \sum_{i=0}^{N-1} x[n]$.
- T is the sum of independent Gaussian random variables so it is a Gaussian RV with mean $N\theta$ and variance $N\sigma^2$.
- Next, we need the joint distribution of $\boldsymbol{x}$ and T. Note that $\boldsymbol{x}$ and T are linked.
- Let $T_0(\boldsymbol{x}) = \sum_{i=0}^{N-1} x[n]$. Then $\mathrm{Prob}(X = \boldsymbol{x}, T! = T_0(\boldsymbol{x})) = 0$, and $\mathrm{Prob}(X = \boldsymbol{x}, T = T_0(\boldsymbol{x})) = p(\boldsymbol{x}; \theta)$.
- Therefore a functional form is $p(\boldsymbol{x}, T; \theta) = p(\boldsymbol{x}; \theta)\delta(T - T_0(\boldsymbol{x}))$. (continued on next slide)

Normal pdf - Hard Version Cont.

- The joint pdf is:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-\theta)^2\right)\delta(T - T_0(\boldsymbol{x})) =$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]^2) - 2\theta T_0(\boldsymbol{x}) + N\theta^2\right)\delta(T - T_0(\boldsymbol{x}))$$

- And the conditional pdf is:

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]^2) - 2\theta T_0(\boldsymbol{x}) + N\theta^2\right)\delta(T - T_0(\boldsymbol{x}))}{\frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{1}{2N\sigma^2}(T - N\theta)^2\right)}$$

- Now we have to simplify to see that the terms with θ cancel out.
- This was much more complicated than the factorization theorem - often times figuring out a sufficient statistic from this method is impossible.

Two Jointly Sufficient Statistics

- Given a known amplitude A, frequency f, and white noise $w[n] \sim N(0, \sigma^2)$ consider the model:

$$x[n] = A\cos(2\pi fn + \theta) + w[n]$$

- What is the distribution of $x[n]$?

$$x[n] \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x[n] - A\cos(2\pi fn + \theta))^2}{\sigma^2} \right)$$

- What is the joint distribution of $\boldsymbol{x}$?

$$\boldsymbol{x} \sim \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp\left(\sum \frac{(x[n] - A\cos(2\pi fn + \theta))^2}{\sigma^2} \right)$$

- Let's use the factorization theorem to find sufficient statistics that can be used to estimate θ. Hint: cos$(a + b) =$ cos(a)cos(b)-sin(a)sin(b)

Two Jointly Sufficient Statistics Cont.

$$x = \frac{1}{\sqrt{2\pi\sigma^2}^{N}} \exp\left(\frac{\sum (x[n] - A\cos(2\pi f n + \theta))^2}{\sigma^2}\right) =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}^{N}} \exp\left(\frac{1}{\sigma^2}\sum\left(x^2[n] - \underbrace{2A\cos(2\pi f n + \theta)\,x[n]}_{\parallel} + A^2\cos^2(2\pi f n + \theta)\right)\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{1}{\sigma^2}\sum x^2[n] \overbrace{- 2A\,x[n]\cos(2\pi f n)\cos(\theta) + 2A\,x[n]\sin(2\pi f n)\sin(\theta)} + A^2\cos^2(2\pi f n + \theta)\right)$$

$$= \underbrace{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{1}{\sigma^2}\sum x^2[n]\right)}_{h(x)} \underbrace{\exp\left(-2A\overbrace{x[n]\cos(2\pi f n)}^{T_1}\cos(\theta) + 2A\overbrace{x[n]\sin(2\pi f n)}^{T_2}\sin\theta + A^2\cos^2(2\pi f n + \theta)\right)}_{g(T_1(x),\,T_2(x),\,\theta)}$$

What are $T_1(\boldsymbol{x}) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f n)$ and $T_2(X) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f n)$?

Minimal Sufficient Statistics

- Minimal sufficient statistic: A sufficient statistic $M(\boldsymbol{x})$ is minimally sufficient if any other sufficient statistic, T, can be expressed as a function $f(M(\boldsymbol{x}))$.
- Intuitively, they take up the least memory compared to any other sufficient statistic.
- M is minimally sufficient if the following statement holds for all possible $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:

$$\frac{p(\boldsymbol{x}_1;\theta)}{p(\boldsymbol{x}_2;\theta)} \text{ is not a function of } \theta \iff M(\boldsymbol{x}_1) = M(\boldsymbol{x}_2)$$

- For most "nice" probability distributions, there is at least one minimal sufficient statistic.

Example of Minimal Sufficient Statistics

- Let Y be a binomial random variable with known size N and unknown probability p.
- Let X be a random variable with a binomial conditional density with size y and unknown probability q, i.e. $f_{X|Y}(X|Y=y) \sim \text{Bin}(y,q)$
- Find a minimal sufficient statistic for (p,q).
- First set up the probability distribution:

$$f(X,Y;p,q) = f_Y(Y;p) f_{X|Y}(X|Y;q) =$$

$$\binom{N}{y}^y (1-)^{N-y} \binom{y}{x} q^x (1-q)^{y-x}$$

$$\text{for } x \in \{0,1,...y\}, y \in \{0,1,...N\}$$

- Let $T_1 = x$ and $T_2 = y$, $h(x,y) = \binom{N}{y}\binom{y}{x}$ and $g(x,y;p,q) = p^y(1-p)^{N-y} q^x (1-q)^{y-x}$. By the factorization theorem (x,y) are sufficient statistics.

Example of Minimal Sufficient Statistics Cont.

- Let (x_1, y_1) and (x_2, y_2) be two samples from $f(X, Y; p, q)$.
- First show that $(T_1(x_1) = x_1 = x_2 = T_2(x_2) \text{ and } T_2(y_1) = y_1 = y_2 = T_2(y_2)) \rightarrow (\frac{p(\boldsymbol{x}_1;\theta)}{p(\boldsymbol{x}_2;\theta)} \text{ is not a function of } \theta)$:

$$\frac{f(x_1, y_1; p, q)}{f(x_2, y_2; p, q)} = \frac{\binom{N}{y_1} p^{y_1} (1-p)^{N-y_1} \binom{y_1}{x_1} q^{x_1} (1-q)^{y_1-x_1}}{\binom{N}{y_2} p^{y_2} (1-p)^{N-y_2} \binom{y_2}{x_2} q^{x_2} (1-q)^{y_2-x_2}} = 1,$$

  which is not a function of p or q.
- In fact, we can see above that to cancel out p and q from the equation, $x_1 = x_2$ and $y_1 = y_2$.
- Therefore, (x, y) are minimal sufficient statistics of p an q.

Why Bother with Sufficient Statistics?

- Sufficient statistics save on memory, which is useful for hardware/microchip applications or when dealing with big data.
- In future lectures, we will be optimizing $p(\boldsymbol{x};\theta) = h(\boldsymbol{x})g(T(\boldsymbol{x}),\theta)$ with respect to θ. The factorization theorem allows us to remove $h(\boldsymbol{x})$.
- There are many performance results related to sufficient statistics, for example when using certain types of probability distributions we can use the sufficient statistic to find the best estimator.

“Something Something Statistics, Something Something Death Star, ... Something Something COMPLETE”

- Let $\boldsymbol{x} = [x[0], ..., x[N-1]]^T$ be a set of samples from a probability distribution $p(\boldsymbol{x}; \theta)$ with unknown parameter θ.
- A sufficient statistic $T(\boldsymbol{x})$ is Complete if for all functions g the following statement holds:

$$\mathrm{E}[g(T); \boldsymbol{\theta}] = 0 \implies p(g(T(\boldsymbol{x})) = 0; \boldsymbol{\theta}) = 1 \quad \forall \boldsymbol{\theta}$$

- This condition ensures that for different values of θ, the probability distributions in the model are different from each other.
- For example, normal RV’s with different means are distinct from each other, and $T(\boldsymbol{x}) = \sum_{n=0}^{N-1} x[n]$ is complete and sufficient to estimate the means.

Example: Bernoulli Trials

- Let $\boldsymbol{x}$ be i.i.d. samples of a bernoulli distribution with probability $\theta \in (0, 1)$. So $x[n] \in \{0, 1\}$
- The probability distribution of individual $x[n]$ is $p(x[n]; \theta) = \theta^{x[n]}(1-\theta)^{1-x[n]}$
- Then the joint probability distribution is found by multiplying the individual distributions:

$$p(\boldsymbol{x}; \theta) = \theta^{\sum_{n=0}^{N-1} x[n]}(1-\theta)^{N-\sum_{n=0}^{N-1} x[n]}$$

- From the factorization theorem, $T(\boldsymbol{x}) = \sum_{n=0}^{N-1} x[n]$ is a sufficient statistic.
- Is it complete?

Example: Bernoulli Trials Cont.

- What is the distribution of a sum of Bernoulli random variables?
- The distribution of T is:

$$p(T(\boldsymbol{x});\theta) = \binom{N}{T}\theta^T(1-\theta)^{N-T}$$

- Check the condition for completeness by assuming $\mathrm{E}[g(T);\theta] = 0$.
- So T is a complete statistic. Why is that good?

Example: Bernoulli Trials Cont.

$$E\, g(T(x)) = \sum_{t=0}^{n} g(t) \binom{N}{t} \theta^{T} (1-\theta)^{N-T} \overset{!}{=} 0$$

Note that if $\theta \in (0,1)$ then:

$$\sum_{t=0}^{n} g(t) \underbrace{\binom{N}{t}}_{>0} \underbrace{\theta^{T}}_{>0} \underbrace{(1-\theta)^{N-T}}_{>0} = 0$$

$$\iff g(t) = 0 \quad \forall t$$

$$\Rightarrow P\left(g(t(x)) = 0, \theta\right) = 1, \quad \forall \theta \; \blacksquare$$

Using Complete Statistics as Estimators

- Let $\hat{\theta}$ be an unbiased estimator you have found.
- Let $T(\boldsymbol{x})$ be a sufficient statistic.

Rao-Blackwell-Lehmann-Scheffe Theorem

The new estimator $\hat{\theta}_T = \mathrm{E}[\hat{\theta}(\boldsymbol{x})|T(\boldsymbol{x})]$ is also an unbiased estimator of θ, but $\mathrm{var}(\hat{\theta}_T) \leq \mathrm{var}(\hat{\theta})$.
If T is complete, $\hat{\theta}_T$ is the minimum variance unbiased (MVU) estimator.

- For a proof, see appendix 5A in Kay-1.
- In the case where we can find a complete sufficient statistic that is an unbiased estimator (e.g., $\frac{1}{N}\sum_{n=0}^{N-1} x[n]$ for Gaussian means), we know that it is the MVU estimator.