X Lecture 8, Oct. 27, 2010

Estimation of non-random parameters

As in the case of random parameter estimation assume that our observation $Y \in \Gamma$ has a conditional distribution P_θ where $\theta \in \Lambda$. However, in this case θ is a fixed value but unknown to us. That is, we either do not know about its statistical properties or it is non-random (fixed) but unknown. In this case, as in the case of random parameters the conditional risk is

$$R_\theta(\hat{\theta}) = E_\theta\{C(\hat{\theta}(y), \theta)\}, \quad \theta \in \Lambda$$

and in the case of squared-error:

$$R_\theta(\hat{\theta}) = E_\theta\{(\hat{\theta}(Y) - \theta)^2\}, \quad \theta \in \Lambda$$

Since we do not have any prior on Λ, we cannot average over θ and we can only average over Y. So, we can only minimize $R_\theta(\hat{\theta})$. But, it is not usually possible to minimize this uniformly for all θ. For example, in the case of mean-squared-error, we can minimize $R_\theta(\hat{\theta}) = E_\theta\{(\hat{\theta}(y)-\theta)^2\}$ for each value of θ, say θ_0, by equating $\hat{\theta}(y)$ with θ_0 and making $R_{\theta_0}(\hat{\theta}) = 0$. But this is not a good solution if θ_0 is not close to the actual value of θ.

So, it is clear that in this case, Minimum Mean-Squared-Error Criterion by itself is not sufficient for finding the estimator for <u>non-random</u> parameters. It is necessary, therefore, to put restrictions on the estimator in order to avoid solutions such as $\hat{\theta}(y) = \theta_0$.

<u>Un-biased estimator</u>

One reasonable restriction is to require that the expected value of the estimate be equal to the actual value of the parameter, i.e.,

$$E_\theta\{\hat{\theta}(y)\} = \theta$$

Such an estimate is called <u>un-biased</u> estimate.

Mean-Squared-Error for an un-biased estimate will be the variance of the estimate. So, the <u>minimum-mean-squared error</u> estimator in this case will be <u>Minimum-Variance</u> <u>Un-biased</u> Estimator (MVUE).

<u>Sufficient Statistics</u>:

Let Δ be an arbitrary set an D be its event class then the function

$T : (\Gamma, G) \rightarrow (\Delta, D)$ is called

a <u>sufficient</u> <u>statistics</u> if the distribution of Y given $T(Y)$ does not depend on θ. It means that knowing $T(Y)$, we do not need to look more into Y to get any clue about θ.

<u>Minimum Sufficient Statistics</u>:

A function $T(Y)$ on (Γ, G) is called the <u>minimal</u> <u>sufficient</u> <u>statistics</u> for $\{P_\theta, \theta \in \Lambda\}$ if it is a function of every other sufficient statistics for $\{P_\theta, \theta \in \Lambda\}$.

This means that the minimum sufficient statistics is the most compact form that the observation can be compressed without destroying the information about θ.

<u>The factorization Theorem</u>:

Let $\{P_\theta; \theta \in \Lambda\}$ have a family of densities $\{p(y|\theta); \theta \in \Lambda\}$. A statistics T is sufficient for θ if and only if there are functions g_θ such that

$$p(y|\theta) = g[T(y)|\theta]h(y) = g_\theta(T(y))h(y)$$

for all $y \in \Gamma$ and $\theta \in \Lambda$.

<u>Proof</u>: Consider the case of discrete Γ.

Suppose Γ is discrete and $\{p(y|\theta); \theta \in \Lambda\}$ satisfies $p(y|\theta) = g_\theta(T(y))h(y)$. Let $p_\theta(y|t) = p(y|T(y)=t)$ be the conditional density of Y given $T(Y)=t$, when $Y \sim P_\theta$.

Then, using Bayes formula:

$$p(y \mid T(y)=t, \theta) = \frac{P(Y=y \mid \theta)\, P[T(Y)=t \mid Y=y, \theta]}{P[T(Y)=t \mid \theta]}$$

But

$$P[T(Y)=t \mid Y=y, \theta] = \begin{cases} 1 & \text{if } T(y)=t \\ 0 & \text{if } T(y) \neq t \end{cases}$$

and $P(Y=y \mid \theta) = p(y \mid \theta)$.

So:

$$p(y \mid T(y)=t, \theta) = \begin{cases} \dfrac{p(y \mid \theta)}{P(T(Y)=t \mid \theta)} & \text{if } T(y)=t \\ 0 & \text{if } T(y) \neq t \end{cases}$$

But, we have

$$P(T(Y)=t \mid \theta) = \sum_{\{y \mid T(y)=t\}} p(y \mid \theta).$$

So,

$$P(T(Y)=t \mid \theta) = \sum_{\{y \mid T(y)=t\}} g(T(y) \mid \theta)\, h(y)$$

$$= g(t \mid \theta) \sum_{\{y \mid T(y)=t\}} h(y)$$

We also have,

$$p(y|\theta) = g(t|\theta)h(y)$$

So:

$$p(y|T(y)=t,\theta) = \begin{cases} \dfrac{h(y)}{\sum_{\{y|T(y)=t\}} h(y)} & \text{if } T(y)=t \\ 0 & \text{if } T(y)\neq t \end{cases}$$

It is seen that this conditional density does not depend on θ. So, $T(y)$ is sufficient statistics for θ.

On the other hand (to prove only if part):

$$p(y|\theta) = p(y|T(y),\theta)\,p(T(y)=T(y)|\theta)$$

Since T is sufficient for θ, $P(y|T(y),\theta)$ depends only on y and not on θ. Also $P[T(y)=T(y)|\theta]$ is a function of $T(y)$ and θ only. Let

$$h(y) \triangleq p(y|T(y),\theta)$$

and

$$g_\theta(T(y)) \triangleq P(T(y)=T(y)|\theta)$$

Then we have

$$p(y|\theta) = g_\theta(T(y))\,h(y)$$

Sufficient Statistics for Hypothesis's Testing

You noticed in previous lectures when we dealt with detection problem, any sort of hypothesis testing ended up with a ratio test dealing with the ratio of $p(y|H_1)$ over $p(y|H_0)$, i.e.,

$$L(y) = \frac{p(y|H_1)}{p(y|H_0)} = \frac{p(y|\theta=1)}{p(y|\theta=0)}$$

Note that,

$$p(y|\theta) = \begin{cases} p(y|\theta=0) & \text{if } \theta=0 \\ \dfrac{p(y|\theta=1)}{p(y|\theta=0)}\, p(y|\theta=0) & \text{if } \theta=1 \end{cases}$$

So, if we choose $h(y) = p(y|\theta=0)$

$$g_\theta(t) = \begin{cases} 1 & \text{if } \theta=0 \\ t & \text{if } \theta=1 \end{cases}$$

Then

$$p(y|\theta) = g_\theta[L(y)]\,h(y)$$

So, $L(y) = \dfrac{p(y|H_1)}{p(y|H_0)}$ is sufficient statistics for H_j.

The Rao-Blackwell Theorem

Suppose $\hat{g}(y)$ is an unbiased estimate of $g(\theta)$ and that T is sufficient for θ.

Define

$$\tilde{g}[T(Y)] = E_\theta\{\hat{g}(Y) \mid T(Y) = T(y)\}$$

Then $\tilde{g}[T(Y)]$ is also an unbiased estimate of $g(\theta)$. Furthermore,

$$Var_\theta(\tilde{g} \mid T(Y)) \leq Var_\theta[\hat{g}(Y)]$$

with equality if and only if,

$$P[\hat{g}(Y) = \tilde{g}(T(Y)) \mid \theta] = 1.$$

Proof: See the text, page 161.

This theorem gives us a means to find MVUE for a parameter starting from any estimate. Of course, in case there is a unique unbiased estimate, that estimate is itself MVUE.

Definition:

The family of distributions $\{P_\theta ; \theta \in \Lambda\}$ is said to be complete if

$$E_\theta\{f(Y)\} = 0 \quad \text{for all } \theta \in \Lambda$$

implies that

$$P[f(Y) = 0 \mid \theta] = 1 \quad \text{for all } \theta \in \Lambda.$$

A sufficient statistics T is said to be complete if its distribution $\{Q_\theta ; \theta \in \Lambda\}$ is complete.

Assume that T is sufficient statistics for θ and $\tilde{g}(T(Y))$ and $g^*(T(Y))$ are functions of $T(Y)$ that are unbiased estimates of $g(\theta)$.

We have

$$E_\theta\{\tilde{g}(T(Y)) - g^*(T(Y))\} = E_\theta\{\tilde{g}(T(Y))\} - E_\theta\{g^*(T(Y))\}$$
$$= g(\theta) - g(\theta) = 0$$

for all $\theta \in \Lambda$

Because of Completeness, we have

$$P[\tilde{g}(T(Y)) = g^*(T(Y))] = 1 \quad \text{all } \theta \in \Lambda.$$

This means that for a complete sufficient statistics the MVUE is unique.

So, the procedure for seeking MVUE can be summarized as:

1) Find a complete sufficient statistics T for $\{P_\theta ; \theta \in \Lambda\}$
2) Find <u>any</u> unbiased estimator $\hat{g}(Y)$ for $\{P_\theta ; \theta \in \Lambda\}$
3) Then find
$$\tilde{g}[T(y)] = E_\theta\{\hat{g}(Y) | T(y) = T(y)\}$$
This is the MVUE for $g(\theta)$.

Maximum - Likelihood Estimation

The above discussed technique is not always applicable either because of complexity or lack of complete sufficient statistics. An alternative is to use Maximum-Likelihood- Estimation or ML estimation. Similar to the case of detection ML technique is optimal only if the parameter to be estimated is uniformly distributed. In such a case

ML detection (estimation as the case is here) is same as the MAP detection (or estimation).

Note that MAP estimation consists in finding θ that maximizes $w(\theta)p(y|\theta)$, i.e.,

$$\hat{\theta}_{MAP}(y) = \arg\left\{\max_{\theta\in\Lambda} w(\theta)p(y|\theta)\right\}$$

In the absence of any prior for θ, we can assume that θ is uniformly distributed, i.e., $w(\theta) = k \quad \forall \theta \in \Lambda$ where $k = \frac{1}{Vol(\Lambda)}$.
For example in case of a phase variable, we may assume $w(\theta) = \frac{1}{2\pi} \quad \theta \in [0, 2\pi]$.
With constant $w(\theta)$, the maximization is done on $p(y|\theta)$, i.e., on the likelihood function of y (given θ):

$$\hat{\theta}_{ML}(y) = \arg\left\{\max_{\theta\in\Lambda} p(y|\theta)\right\}$$

maximizing $p(y|\theta)$ is equivalent to maximizing $\log p(y|\theta)$. If $\log p(y|\theta)$

is smooth enough, we have

$$\frac{\partial}{\partial \theta} \log p_1(y|\theta) \Big|_{\theta = \hat{\theta}_{ML}(y)} = 0$$

Note that in case of MAP estimation, we needed to maximize $p(\theta|y)$, i.e.,

$$p(\theta|y) = \frac{w(\theta)\, p(y|\theta)}{p(y)}$$

or its log: $\log\{p(\theta|y)\}$. So, for MAP, we had to set

$$\left[\frac{\partial \log p(y|\theta)}{\partial \theta} + \frac{\partial \log w(\theta)}{\partial \theta}\right]_{\theta = \hat{\theta}_{MAP}} = 0$$

in the case of non-random parameters, or, equivalently, uniformly distributed parameters the second term vanishes and we have ML estimation.

<u>Example</u>: Consider the problem of estimating a parameter θ in additive noise, i.e.,

$$Y_i = \theta + N_i \qquad i=1,2,\dots,n$$

where $N_i \sim (0, \sigma_N^2)$. i.i.d

1) First assume that $\theta \sim N(0, \sigma_\theta^2)$. Find MAP estimate.
2) Assume that θ is a non-random parameter and find ML estimate.

$$P(\theta|\underline{y}) = \frac{w(\theta)p(\underline{y}|\theta)}{\int_\Lambda w(\theta)p(\underline{y}|\theta)d\theta} = \frac{w(\theta)p(\underline{y}|\theta)}{p(\underline{y})}$$

$$p(\underline{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_N} e^{\frac{-(y_i-\theta)^2}{2\sigma_N^2}}$$

and

$$w(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\theta} e^{-\frac{(\theta)^2}{2\sigma_\theta^2}}$$

We do not need to find $p(\underline{y})$ since it is not dependant on θ. We can just maximize $w(\theta)p(\underline{y}|\theta)$.

$$p(\theta|\underline{y}) = K'(\underline{y}) \exp\left[-\frac{1}{2\sigma_m^2}\left(\theta - \frac{\sigma_m^2}{\sigma_N^2}\left(\sum_{i=1}^{n} y_i\right)\right)^2\right]$$

where

$$\sigma_m^2 = \frac{\sigma_\theta^2 \sigma_N^2}{n\sigma_\theta^2 + \sigma_N^2}$$

The MAP estimate results by choosing

$$\hat{\theta}_{MAP}(\underline{y}) = \frac{\sigma_m^2}{\sigma_N^2} \sum_{i=1}^{n} y_i$$

$$= \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_N^2/n} \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

Note that $T(\underline{y}) = \sum_{i=1}^{n} y_i$ is a sufficient statistics for θ.

2) Now for the case where θ is non-random.

Since θ has no prior, we take $w(\theta) = \text{Constant}$ and find ML estimate by taking derivative of $\log p(\underline{y}|\theta)$, i.e.,

$$\frac{\partial}{\partial\theta} \log p(\underline{y}|\theta) = \frac{\partial}{\partial\theta} \log\left\{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left[-\frac{(y_i-\theta)^2}{2\sigma_N^2}\right]\right\}$$

$$= \frac{1}{\sigma_N^2}\left(\sum_{i=1}^{n} y_i - n\theta\right) = 0$$

or

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

That is, the ML estimate is the sample mean of observations.

Example 2: Consider estimating $g(\theta)$ which is a non-linear function of θ in Gaussian Noise, i.e.,

$$y_i = g(\theta) + N_i \qquad i=1,\cdots,n$$

where $\quad N_i \sim (0, \sigma_N^2) \quad$ i.i.d.

$$p(\underline{y}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_N} \exp\left[-\frac{1}{2\sigma_N^2}\sum_{i=1}^{n}(y_i - g(\theta))^2\right]$$

Here,

$$\frac{\partial}{\partial\theta}\log p(\underline{y}|\theta) = \frac{1}{\sigma_N^2}\sum_{i=1}^{n}[y_i - g(\theta)]\frac{\partial g(\theta)}{\partial\theta}\bigg|_{\theta=\hat{\theta}_{ML}} = 0$$

or

$$\left[\frac{1}{\sigma_N^2}\frac{\partial g(\theta)}{\partial\theta}\right]\left[\frac{1}{n}\sum_{i=1}^{n}y_i - g(\theta)\right]\bigg|_{\theta=\hat{\theta}_{ML}} = 0$$

So

$$g(\hat{\theta}_{ML}) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

or

$$\hat{\theta}_{ML}(\underline{y}) = g^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]$$