Lecture 7, Oct. 20, 2010

In detection problems, we have to choose between two (or at most, a finite number of) hypotheses. For example, we may want to know whether a target is there or not, or whether an object is moving or stationary.

In estimation problems, we may need to choose from a continuum of possible values.

Estimation problems can be modelled similar to detection problem.

- The random variable Y is observed: $Y \in \Gamma$
- $Y \sim P_\theta$, i.e., Y is distributed according to P_θ depending on $\theta \in \Lambda$. Λ is called the parameter set. P_θ is a distribution on the observation space $(\Gamma, \mathcal{G})$.

That is given θ, Y has probability density function (pdf): $p(y|\theta)$.

- The goal of the estimation problem is to find an estimate of θ, say, $\hat{\theta}$, based

on the observation of y. That is, to find a mapping: $\hat{\theta}: \Gamma \to \Lambda$

where $\hat{\theta}(y)$ is the "best" match for the true θ, according to certain criterion (<u>cost function</u>).

A cost function $C(\cdot,\cdot)$ is a (usually, non-nega function from $\Lambda \times \Lambda$ to R such that $C(\hat{\theta}, \theta)$ is the cost of deciding $\hat{\theta}$ when the actual value of the parameter is θ.

The conditional risk (or average cost) is

$$R_\theta(\hat{\theta}) = E_\theta \{ C(\hat{\theta}(Y), \theta) \} \qquad \forall \theta \in \Lambda$$

$R_\theta(\hat{\theta})$ is the risk of deciding $\hat{\theta}$ averaged over all Y for a given θ.

The average or Bayes risk is the average of $R_\theta(\hat{\theta})$ over all $\theta \in \Lambda$, i.e.,

$$r(\hat{\theta}) = E[R_\theta(\hat{\theta})]$$

we have

$$r(\hat{\theta}) = E[C(\hat{\theta}(Y), \theta)]$$

$$= E[E\{C(\hat{\theta}(Y), \theta) | Y\}]$$

From the above, one can conclude that, the Bayes estimator for θ can be found by minimizing the cost for each $y \in \Gamma$, i.e., minimizing,

$$E\{C[\hat{\theta}(y), \theta] | Y = y\}$$

Denoting the conditional density of θ (given y) by $w(\theta|y)$, we need to minimize,

$$\int_{\Lambda} C[\hat{\theta}(y), \theta] w(\theta|y) \, d\theta .$$

Example:

- Minimum-Mean-Squared-Error (MMSE):

A very common cost function is the square of the difference between the actual value θ and its estimate $\hat{\theta}$, i.e.,

$$C[\hat{\theta}, \theta] = (\hat{\theta} - \theta)^2 \qquad (\hat{\theta}, \theta) \in \mathbb{R}^2$$

The Bayes risk, in this case, is

$$E[(\hat{\theta}(Y) - \theta)^2]$$

called the mean-squared-error.

The posterior cost given $Y=y$, is

$$E[(\hat{\theta}(y)-\theta)^2 | Y=y] = E\{[\hat{\theta}(y)]^2 | Y=y\}$$

$$- 2E[\hat{\theta}(y)\,\theta | Y=y]$$

$$+ E\{\theta^2 | Y=y\}$$

$$= [\hat{\theta}(y)]^2 - 2\hat{\theta}(y) E[\theta | Y=y]$$

$$+ E[\theta^2 | Y=y]$$

Taking the derivative and equating it to zero, we get:

$$\hat{\theta}_{MMSE}(y) = E[\theta | Y=y]$$

That is, the MMSE estimate is the conditional mean of the parameter value given the observation y

Example: Estimation of Parameters of an Exponential Distribution.

Let $\Lambda = (0, \infty)$ and $\Gamma = \mathbb{R}$ and

$$p(y|\theta) = \begin{cases} \theta e^{-\theta y} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0 \end{cases}$$

That is, y has <u>exponential</u> <u>density</u> with parameter θ.

Exponential density models inter-arrival time of Markov arrivals, e.g., in packet and circuit switched networks, or the time between failures for components. θ is the rate of failure (or arrival).
We also assume that the parameter θ has an exponential distribution $w(\theta)$:

$$w(\theta) = \begin{cases} \alpha e^{-\alpha\theta} & \theta \geq 0 \\ 0 & \theta < 0 \end{cases}$$

with an $\alpha > 0$ parameter which is a priori known.

Based on $w(\theta)$ and $p(y|\theta)$ we can find the <u>a posteriori</u> for θ given y,

$$w(\theta|y) = \frac{\alpha\theta e^{-(\alpha+y)\theta}}{\int_0^\infty \alpha\theta e^{-(\alpha+y)\theta} d\theta}$$

$$= (\alpha+y)^2 \theta e^{-\theta(\alpha+y)}$$

for all $\theta \geq 0$ and $y \geq 0$

$$w(\theta|y) = 0 \quad \text{for } \theta < 0 \text{ or } y < 0.$$

The MMSE estimate is:

$$\hat{\theta}_{MMSE} = E[\theta|y] = \int_0^\infty \theta w(\theta|y)\, d\theta$$

$$= (\alpha+y)^2 \int_0^\infty \theta^2 e^{-\theta(\alpha+y)}\, d\theta$$

$$= \frac{2}{\alpha+y}$$

We see that the estimate is in inverse proportion of y. This can be intuitively appealing as large interval between arrival (failures) indicates low rate of arrival (failure).

The Bayes risk or MMSE error is:

$$MMSE = r(\hat{\theta}_{MMSE}) = E[E\{\hat{\theta}_{MMSE}(Y) - \theta\}^2 | Y]$$

$$= E[E[\theta - E(\theta|Y)]^2 | Y]$$

$$= E[Var(\theta|Y)]$$

The minimum MSE for a given y is:

$$Var(\theta|Y=y) = E[\theta^2|Y=y] - E^2[\theta|Y=y]$$

$$= \int_0^\infty \theta^2 w(\theta|y)\, d\theta - [\hat{\theta}_{MMSE}(y)]^2$$

$$Var(\theta|Y=y) = (\alpha+y)^2 \int_0^\infty \theta^3 e^{-\theta(\alpha+y)} d\theta$$

$$- \frac{4}{(\alpha+y)^2}$$

$$= \frac{2}{(\alpha+y)^2}$$

So,

$$mmse = E\left\{\frac{2}{(\alpha+y)^2}\right\} = \int_0^\infty \frac{2}{(\alpha+y)^2} p(y)\, dy$$

where

$$p(y) = \int_0^\infty w(\theta)\, p(y|\theta)\, d\theta$$

$$= \int_0^\infty \alpha\theta e^{-(\alpha+y)\theta} d\theta = \frac{\alpha}{(\alpha+y)^2}$$

therefore,

$$mmse = \int_0^\infty \frac{2\alpha}{(\alpha+y)^4} dy = \boxed{\frac{2}{3\alpha^2}}$$

Estimation of Amplitude

Let $\Gamma = \mathbb{R}^n$ and $\Lambda = \mathbb{R}$ and

$$Y_k = N_k + \theta s_k \qquad k = 1, \dots, n$$

where $\underline{N} \sim N(0, \Sigma)$, $\underline{s} = (s_1, \dots, s_n)$ is a known signal and $\theta \sim N(\mu, \nu^2)$.

Furthermore, assume that $\underline{N}$ and θ are independent.

Given $\theta = \theta$, we have $\underline{Y} \sim N(\theta \underline{s}, \Sigma)$.

The posterior density is:

$$w(\theta | \underline{y}) = \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underline{y} - \theta \underline{s})^T \Sigma^{-1} (\underline{y} - \theta \underline{s})\right] \frac{1}{\sqrt{2\pi}\nu} e^{-\frac{(\theta-\mu)^2}{2\nu^2}}}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underline{y} - \theta \underline{s})^T \Sigma^{-1} (\underline{y} - \theta \underline{s})\right] \frac{1}{\sqrt{2\pi}\nu} e^{-\frac{(\theta-\mu)^2}{2\nu^2}} d\theta}$$

or

$$w(\theta | \underline{y}) = K(\underline{y}) \exp\left\{-\frac{\theta^2}{2}\left(d^2 + \frac{1}{\nu^2}\right) + \theta\left(s^T \Sigma^{-1} \underline{y} + \frac{\mu}{\nu^2}\right)\right.$$

where $d^2 = \underline{s}^T \Sigma^{-1} \underline{s}$ and $K(\underline{y})$ is a factor that is only a function of $\underline{y}$ and does not depend on θ.

Since $w(\theta | \underline{y})$ is an exponential function of a quadratic term in θ, it is Gaussian

and can be written as:

$$w(\theta|\underline{y}) = \frac{1}{\sqrt{2\pi}\,q} e^{-(\theta-m)^2/2q^2}$$

$$= \frac{e^{-m^2/2q^2}}{\sqrt{2\pi}\,q} e^{-\frac{\theta^2}{2q^2} + \theta\frac{m}{q^2}}$$

Comparing this with the former expression for $w(\theta|\underline{y})$, we see that:

$$q^2 = \left(d^2 + \frac{1}{v^2}\right)^{-1}$$

and

$$m = \frac{\left(\underline{s}^T\Sigma^{-1}\underline{y} + \frac{\mu}{v^2}\right)}{\left(d^2 + \frac{1}{v^2}\right)}$$

That is, given $\underline{Y} = \underline{y}$, $\theta \sim N(m, q^2)$ where m and q^2 are given above.

With these values of q^2 and m, we have

$$K(\underline{y}) = \frac{e^{-m^2/2q^2}}{\sqrt{2\pi}\,q}.$$

$\hat{\theta}_{MMSE}(\underline{y})$ is the mean of θ given $\underline{y}$, i.e.,

$$\hat{\theta}_{MMSE} = \frac{\underline{s}^T\Sigma^{-1}\underline{y} + \frac{\mu}{v^2}}{\left(d^2 + \frac{1}{v^2}\right)} = \frac{v^2d^2\hat{\theta}_1(\underline{y}) + \mu}{v^2d^2 + 1}$$

where $\hat{\theta}_1(\underline{y}) = \underline{s}^T\Sigma^{-1}\underline{y}/d^2$.

The minimum-mean-squared-error is:

$$MMSE = E\{Var(\theta|\underline{y})\} = \frac{1}{d^2 + \frac{1}{\nu^2}}$$

$$= \frac{\nu^2}{d^2\nu^2 + 1}$$

Discussion

Looking at the expression for MMSE estimate

$$\hat{\theta}_{MMSE}(\underline{y}) = \frac{\nu^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{\nu^2 d^2 + 1}$$

we see that there are two parts involved $\hat{\theta}_1(\underline{y}) = \underline{s}^T \Sigma^{-1} \underline{y} / d^2$ which depends on the received signal $\underline{y}$ and μ which is our *a priori* knowledge about the parameter to be estimated.

ν^2 is the variance of θ. Large ν^2 indicates less certainty about θ (more spread around μ). But smaller ν^2 means more trust in θ's mean (μ) since small ν^2 means less spread around the mean. Also d^2 (as discussed before) is a

measure of Signal-to-Noise-Ratio.

So small $d^2\nu^2$ is a sign of either more a priori knowledge or a sign of noisy measurement or both. In such a case it is clear that we have to pay less attention to $\underline{y}$ and trust more our prior knowledge. So, it is not surprising that

$$\lim_{d^2\nu^2 \to 0} \frac{\nu^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{\nu^2 d^2 + 1} \to \mu$$

on the other hand if $d^2\nu^2$ is large it means good observation an poor prior knowledge. In this case.

$$\lim_{d^2\nu^2 \to \infty} \frac{\nu^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{\nu^2 d^2 + 1} \to \hat{\theta}_1(\underline{y})$$

That is, in this case we ignore μ and decide based on the received signal.

A special case is when the noise is i.i.d., i.e., $\Sigma = \sigma^2 I$ and $\underline{s}$ is constant, say $\underline{s} = (1, 1, \ldots, 1)$.

Then $\nu^2 d^2 = \frac{n\nu^2}{\sigma^2}$ and

$$\hat{\theta}_1(\underline{y}) = \bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$$

in this case, in the absence of observation we estimate θ as μ, but as we take more samples we will have more confidence on the sample mean, $\bar{y}$. At $n \to \infty$, we disregard the prior mean (μ completely. The convergence speed depends on the ratio $\frac{\nu^2}{\sigma^2}$.

Joint estimation:

Assume that instead of a scalar (single) parameter θ, we deal with $\underline{\theta} = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ for example $\underline{\theta} = (\phi, f, A)$ where ϕ, f and A are phase, frequency and amplitude or $\underline{\theta} = (f_1, f_2, f_3)$ where f_i, $i = 1, 2, 3$ is the i-th harmonic of a given signal.

In this case, the cost function is of the form

$$C(\hat{\underline{\theta}}, \underline{\theta}) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$$

A desirable property of a cost function is that it be per-component, i.e.,

$$C(\hat{\underline{\theta}}, \underline{\theta}) = \sum_{i=1}^{m} C_i(\hat{\theta}_i, \theta_i)$$

An example is the Euclidean norm, i.e.,

$$C(\hat{\underline{\theta}}, \underline{\theta}) = \|\hat{\underline{\theta}} - \underline{\theta}\| = \sum_{i=1}^{m} (\hat{\theta}_i - \theta_i)^2$$

another possibility is absolute-value sum:

$$C(\hat{\underline{\theta}}, \underline{\theta}) = \sum_{i=1}^{m} |\hat{\theta}_i - \theta_i|$$

A generalization of the Euclidean norm is a weighted distance defined as:

$$C(\hat{\underline{\theta}}, \underline{\theta}) = (\hat{\underline{\theta}} - \underline{\theta})^T A (\hat{\underline{\theta}} - \underline{\theta})$$

where A is a symmetric, positive-definite matrix. This cost function can be useful when one needs to estimate different parameters with varying accuracy.

We can write the conditional risk for this cost function as:

$$E\left[(\hat{\underline{\theta}}(y)-\underline{\theta})^T A(\hat{\underline{\theta}}(y)-\underline{\theta})\,|\,Y=y\right]$$
$$=[\hat{\underline{\theta}}(y)]^T A\,\hat{\underline{\theta}}(y) - 2[\hat{\underline{\theta}}(y)]^T A\,E[\underline{\theta}\,|Y=y]$$
$$+E\{\underline{\theta}^T A\,\underline{\theta}\,|Y=y\}$$

To minimize this, we take gradient with respect to $\hat{\underline{\theta}}(y)$ and equate it to zero:

$$\nabla_{\hat{\underline{\theta}}(y)} E\left[C(\hat{\underline{\theta}}(y),\underline{\theta}\right] = 2A\,\hat{\underline{\theta}}(y) - 2A\,E[\underline{\theta}\,|Y=y]$$

So, the Bayes estimate is given by:

$$2A\,\hat{\underline{\theta}}_B(y) = 2A\,E\{\underline{\theta}\,|Y=y\}$$

multiplying by A^{-1} (from left) gives:

$$\hat{\underline{\theta}}_B(y) = E\{\underline{\theta}\,|Y=y\}$$

We observe that this estimate does not depend on A. However, the resulting average risk (the Bayes risk) depends on A. It can be shown that

$$r(\hat{\underline{\theta}}_B) = tr\{A\,E[Cov(\underline{\theta}\,|Y)]\}$$

where $tr(\cdot)$ is the trace operator (the sum of diagonal terms of the matrix).

Minimum-Mean-Absolute-Error (MMAE)

An alternative cost function is

$$C[\hat{\theta}, \theta] = |\hat{\theta} - \theta|, \quad (\hat{\theta}, \theta) \in \mathbb{R}^2$$

The Bayes risk, in this case, is

$E[|\hat{\theta}(Y) - \theta|]$ and has to be minimized

This is called minimum-mean-absolute-error (MMAE) estimate.

The conditional risk is:

$$E[|\hat{\theta}(Y) - \theta| \,|\, Y = y] = \int_{-\infty}^{\infty} |\hat{\theta}(y) - \theta| W(\theta|y)\, d\theta$$

$$= \int_{-\infty}^{\hat{\theta}(y)} (\hat{\theta}(y) - \theta) W(\theta|y)\, d\theta + \int_{\hat{\theta}(y)}^{\infty} (-\hat{\theta}(y) + \theta) W(\theta|y)\, d\theta$$

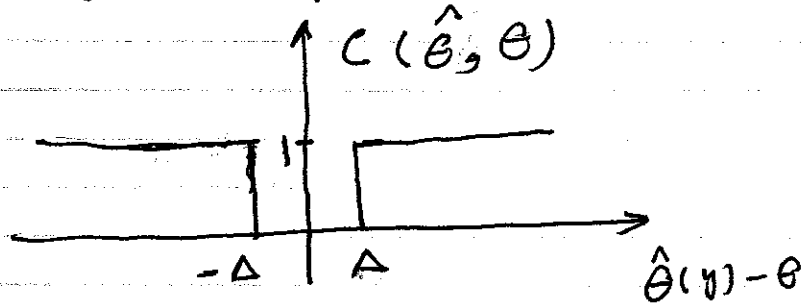
Taking derivative of the above and equating to zero, we get

$$\int_{-\infty}^{\hat{\theta}_{MMAE}(y)} W(\theta|y)\, d\theta = \int_{\hat{\theta}_{MMAE}(y)}^{\infty} W(\theta|y)\, d\theta$$

So, for absolute error, the estimate is the median of the conditional density $W(\theta|y)$.

Uniform Cost function: Maximum A Posteriori Probability (MAP) estimate

Another cost function is the so called uniform cost function:



$$C(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } |\hat{\theta} - \theta| \leq \Delta \\ 1 & \text{if } |\hat{\theta} - \theta| > \Delta \end{cases}$$

The Bayes risk for this cost function is:

$$E[C(\hat{\theta}(y), \theta) | Y = y] = P[|\hat{\theta}(y) - \theta| > \Delta | Y = y]$$

$$= 1 - P[|\hat{\theta}(y) - \theta| \leq \Delta | Y = y]$$

To minimize Bayes risk, we need to maximize

$$P[|\hat{\theta}(y) - \theta| \leq \Delta | Y = y]$$

But,

$$P[|\hat{\theta}(y) - \theta| \le \Delta | Y = y] = \int_{\hat{\theta}(y)-\Delta}^{\hat{\theta}(y)+\Delta} w(\theta|y)\, d\theta$$

When Δ is small the best choice of $\hat{\theta}$ is the point where $w(\theta|y)$ is maximum. That is the reason this is called Maximum A posteriori Probability estimate (MAP).

The three estimates can be found from the conditional a posteriori density

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

The MMSE estimate is the Mean of $w(\theta|y)$

The MMAE estimate is the Median of $w(\theta|y)$

and MAP estimate is the value of θ that maximizes $w(\theta|y)$.