Parameter Estimation

Saravanan Vijayakumaran
sarva@ee.iitb.ac.in

Department of Electrical Engineering
Indian Institute of Technology Bombay

March 20, 2015

Parameter Estimation

- Hypothesis testing was about making a choice between discrete states of nature
- Parameter or point estimation is about choosing from a continuum of possible states

Example

- Consider a manufacturer of clothes for newborn babies
- She wants her clothes to fit at least 50% of newborn babies. Clothes can be loose but not tight. She also wants to minimize material used.
- Since babies are made up of a large number of atoms, their length is a Gaussian random variable (by Central Limit Theorem)

$$\text{Baby Length} \sim \mathcal{N}(\mu, \sigma^2)$$

- Only knowledge of μ is required to achieve her goal of 50% fit
- But μ is unknown and she is interested in estimating it
- What is a good estimator of μ? If she wants her clothes to fit at least 75% of the newborn babies, is knowledge of μ enough?

System Model for Parameter Estimation

- Consider a family of distributions

$$\mathbf{Y} \sim P_{\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \Lambda$$

  where the observation vector $\mathbf{Y} \in \mathbb{R}^n$ and $\Lambda \subseteq \mathbb{R}^m$ is the parameter space. $\boldsymbol{\theta}$ itself can be a realization of a random variable $\boldsymbol{\Theta}$

Example

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

where μ and σ are unknown. Here $\boldsymbol{\theta} = \begin{bmatrix}\mu & \sigma\end{bmatrix}^T$, $\Lambda = \mathbb{R} \times \mathbb{R}^+$.
The parameters μ and σ can themselves be random variables.

- The goal of parameter estimation is to find $\boldsymbol{\theta}$ given $\mathbf{Y}$
- An estimator is a function from the observation space to the parameter space

$$\hat{\boldsymbol{\theta}} : \mathbb{R} \to \Lambda$$

Which is the Optimal Estimator?

- Assume there is a cost function C

$$C : \Lambda \times \Lambda \rightarrow \mathbb{R}$$

such that $C[\mathbf{a}, \boldsymbol{\theta}]$ is the cost of estimating the true value of $\boldsymbol{\theta}$ as $\mathbf{a}$

- Examples of cost functions for scalar θ

Squared Error $C[a, \theta] = (a - \theta)^2$

Absolute Error $C[a, \theta] = |a - \theta|$

Threshold Error $C[a, \theta] = \begin{cases} 0 & \text{if } |a - \theta| \leq \Delta \\ 1 & \text{if } |a - \theta| > \Delta \end{cases}$

Which is the Optimal Estimator?

- Suppose that the parameter $\boldsymbol{\theta}$ is the realization of a random variable $\boldsymbol{\Theta}$
- With an estimator $\hat{\boldsymbol{\theta}}$ we associate a conditional cost or risk conditioned on $\boldsymbol{\theta}$

$$r_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\left\{C\left[\hat{\boldsymbol{\theta}}(\mathbf{Y}), \boldsymbol{\theta}\right]\right\}$$

- The average risk or Bayes risk is given by

$$R(\hat{\boldsymbol{\theta}}) = E\left\{r_{\boldsymbol{\Theta}}(\hat{\boldsymbol{\theta}})\right\}$$

- The optimal estimator is the one which minimizes the Bayes risk

Which is the Optimal Estimator?

- Given that

$$r_\theta(\hat{\theta}) = E_\theta\left\{C\left[\hat{\theta}(\mathbf{Y}), \theta\right]\right\} = E\left\{C\left[\hat{\theta}(\mathbf{Y}), \mathbf{\Theta}\right] \middle| \mathbf{\Theta} = \theta\right\}$$

  the average risk or Bayes risk is given by

$$\begin{aligned} R(\hat{\theta}) &= E\left\{C\left[\hat{\theta}(\mathbf{Y}), \mathbf{\Theta}\right]\right\} \\ &= E\left\{E\left\{C\left[\hat{\theta}(\mathbf{Y}), \mathbf{\Theta}\right] \middle| \mathbf{Y}\right\}\right\} \\ &= \int E\left\{C\left[\hat{\theta}(\mathbf{Y}), \mathbf{\Theta}\right] \middle| \mathbf{Y} = \mathbf{y}\right\} p_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y} \end{aligned}$$

- The optimal estimate for θ can be found by minimizing for each $\mathbf{Y} = \mathbf{y}$ the posterior cost

$$E\left\{C\left[\hat{\theta}(\mathbf{y}), \mathbf{\Theta}\right] \middle| \mathbf{Y} = \mathbf{y}\right\}$$

Minimum-Mean-Squared-Error (MMSE) Estimation

- Consider a scalar parameter θ
- $C[a, \theta] = (a - \theta)^2$
- The posterior cost is given by

$$
\begin{aligned}
E\left\{ (\hat{\theta}(\mathbf{y}) - \Theta)^2 \middle| \mathbf{Y} = \mathbf{y} \right\} &= \left[\hat{\theta}(\mathbf{y})\right]^2 - 2\hat{\theta}(\mathbf{y}) E\left\{ \Theta \middle| \mathbf{Y} = \mathbf{y} \right\} \\
&\quad + E\left\{ \Theta^2 \middle| \mathbf{Y} = \mathbf{y} \right\}
\end{aligned}
$$

- Differentiating posterior cost wrt $\hat{\theta}(\mathbf{y})$, the Bayes estimate is

$$
\hat{\theta}_{MMSE}(\mathbf{y}) = E\left\{ \Theta \middle| \mathbf{Y} = \mathbf{y} \right\}
$$

Example: MMSE Estimation

- Suppose X and Y are jointly Gaussian random variables
- Let the joint pdf be given by

$$p_{XY}(x, y) = \frac{1}{2\pi|\mathbf{C}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{s}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{s}-\boldsymbol{\mu})\right)$$

where $\mathbf{s} = \begin{bmatrix} x \\ y \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$

- Suppose Y is observed and we want to estimate X
- The MMSE estimate of X is

$$\hat{X}_{MMSE}(y) = E\left[X \middle| Y = y\right]$$

- The conditional density of X given $Y = y$ is

$$p(x|y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

Example: MMSE Estimation

- The conditional density of X given $Y = y$ is a Gaussian density with mean

$$\mu_{X|y} = \mu_x + \frac{\sigma_x}{\sigma_y}\rho(y - \mu_y)$$

  and variance

$$\sigma^2_{X|y} = (1 - \rho^2)\sigma^2_x$$

- Thus the MMSE estimate of X given $Y = y$ is

$$\hat{X}_{MMSE}(y) = \mu_x + \frac{\sigma_x}{\sigma_y}\rho(y - \mu_y)$$

Maximum A Posteriori (MAP) Estimation

- In some situations, the conditional mean may be difficult to compute
- An alternative is to use MAP estimation
- The MAP estimator is given by

$$\hat{\boldsymbol{\theta}}_{MAP}(\mathbf{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\,(\boldsymbol{\theta}|\mathbf{y})$$

  where p is the conditional density of $\mathbf{\Theta}$ given $\mathbf{Y}$.

- It can be obtained as the optimal estimator for the threshold cost function

$$C[a,\theta] = \begin{cases} 0 & \text{if } |a-\theta| \leq \Delta \\ 1 & \text{if } |a-\theta| > \Delta \end{cases}$$

  for small $\Delta > 0$

Maximum A Posteriori (MAP) Estimation

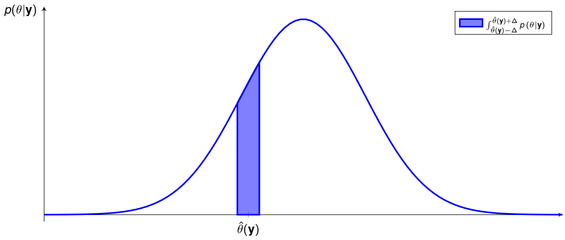
- For the threshold cost function, we have[1]

$$
\begin{aligned}
& E\left\{C\left[\hat{\theta}(\mathbf{y}), \Theta\right] \middle| \mathbf{Y} = \mathbf{y}\right\} \\
&= \int_{-\infty}^{\infty} C[\hat{\theta}(\mathbf{y}), \theta] p(\theta|\mathbf{y}) \; d\theta \\
&= \int_{-\infty}^{\hat{\theta}(\mathbf{y})-\Delta} p(\theta|\mathbf{y}) \; d\theta + \int_{\hat{\theta}(\mathbf{y})+\Delta}^{\infty} p(\theta|\mathbf{y}) \; d\theta \\
&= \int_{-\infty}^{\infty} p(\theta|\mathbf{y}) \; d\theta - \int_{\hat{\theta}(\mathbf{y})-\Delta}^{\hat{\theta}(\mathbf{y})+\Delta} p(\theta|\mathbf{y}) \; d\theta \\
&= 1 - \int_{\hat{\theta}(\mathbf{y})-\Delta}^{\hat{\theta}(\mathbf{y})+\Delta} p(\theta|\mathbf{y}) \; d\theta
\end{aligned}
$$

- The Bayes estimate is obtained by maximizing the integral in the last equality

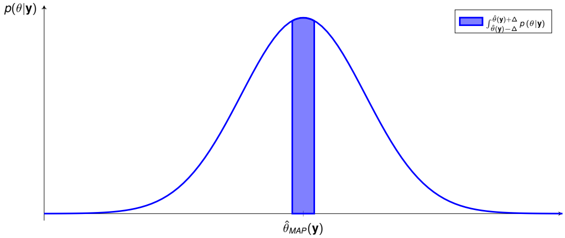
[1] Assume a scalar parameter θ for illustration

Maximum A Posteriori (MAP) Estimation



- The shaded area is the integral $\int_{\hat{\theta}(\mathbf{y})-\Delta}^{\hat{\theta}(\mathbf{y})+\Delta} p(\theta|\mathbf{y}) \; d\theta$
- To maximize this integral, the location of $\hat{\theta}(\mathbf{y})$ should be chosen to be the value of θ which maximizes $p(\theta|\mathbf{y})$

Maximum A Posteriori (MAP) Estimation



- This argument is not airtight as $p(\theta|\mathbf{y})$ may not be symmetric at the maximum
- But the MAP estimator is widely used as it is easier to compute than the MMSE estimator

Maximum Likelihood (ML) Estimation

- The ML estimator is given by

$$\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\mathbf{y}|\boldsymbol{\theta}\right)$$

  where p is the conditional density of $\mathbf{Y}$ given $\boldsymbol{\Theta}$.

- It is the same as the MAP estimator when the prior probability distribution of $\boldsymbol{\Theta}$ is uniform

$$\hat{\boldsymbol{\theta}}_{MAP}(\mathbf{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \frac{p\left(\boldsymbol{\theta}, \mathbf{y}\right)}{p(\mathbf{y})} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \frac{p\left(\mathbf{y}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

- It is also used when the prior distribution is not known

Example 1: ML Estimation

- Suppose we observe $Y_i, \; i = 1, 2, \ldots, M$ such that

$$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

  where Y_i's are independent, μ is unknown and σ^2 is known

- The ML estimate is given by

$$\hat{\mu}_{ML}(\mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} y_i$$

Example 2: ML Estimation

- Suppose we observe $Y_i, \ i = 1, 2, \ldots, M$ such that

  $$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

  where Y_i's are independent, both μ and σ^2 are unknown
- The ML estimates are given by

$$\begin{aligned}\hat{\mu}_{ML}(\mathbf{y}) &= \frac{1}{M}\sum_{i=1}^{M} y_i \\ \hat{\sigma}^2_{ML}(\mathbf{y}) &= \frac{1}{M}\sum_{i=1}^{M} (y_i - \hat{\mu}_{ML}(\mathbf{y}))^2\end{aligned}$$

Example 3: ML Estimation

- Suppose we observe $Y_i, \; i = 1, 2, \ldots, M$ such that

$$Y_i \sim \text{Bernoulli}(p)$$

  where Y_i's are independent and p is unknown

- The ML estimate of p is given by

$$\hat{p}_{ML}(\mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} y_i$$

Example 4: ML Estimation

- Suppose we observe $Y_i, \; i = 1, 2, \ldots, M$ such that

$$Y_i \sim \text{Uniform}[0, \theta]$$

  where Y_i's are independent and θ is unknown
- The ML estimate of θ is given by

$$\hat{\theta}_{ML}(\mathbf{y}) = \max\left(y_1, y_2, \ldots, y_{M-1}, y_M\right)$$

Reference

- Chapter 4, *An Introduction to Signal Detection and Estimation*, H. V. Poor, Second Edition, Springer Verlag, 1994.

Thanks for your attention