ELEG 5633 Detection and Estimation
Maximum Likelihood Estimation

Jingxian Wu

Department of Electrical Engineering
University of Arkansas

Outline

- Classical Estimation
- Maximum Likelihood Estimation
- Asymptotics
- MLE for Transformed Parameters

Classical Estimation

- Classical Estimation: θ is deterministic but unknown.
- The generative (or forward) model under classical setting

$$\theta \to p(x|\theta) \to x$$

  which involves the likelihood only.
- The generative (or forward) model under Bayesian setting

$$p(\theta) \to \theta \to p(x|\theta) \to x$$

  which involves the prior and likelihood.

Basic Concepts

- Loss $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$
- Risk: $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{E}_x[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]$
- Bias: $\mathsf{bias}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_x[\hat{\boldsymbol{\theta}}(x)] - \boldsymbol{\theta}$
- An estimator is unbiased if $\mathsf{bias}(\hat{\boldsymbol{\theta}}) = 0$ for all $\boldsymbol{\theta} \in \Theta$.
- Variance:

$$\begin{aligned} \mathsf{var}(\hat{\boldsymbol{\theta}}) &= \mathsf{tr}\Big(\mathbb{E}\Big[(\hat{\boldsymbol{\theta}}(x) - \mathbb{E}\hat{\boldsymbol{\theta}}(x))(\hat{\boldsymbol{\theta}}(x) - \mathbb{E}\hat{\boldsymbol{\theta}}(x))^T\Big]\Big) \\ &= \mathbb{E}\big[\|\hat{\boldsymbol{\theta}}(x) - \mathbb{E}\hat{\boldsymbol{\theta}}(x)\|_2^2\big] \end{aligned}$$

Mean Square Error (MSE)

$$
\begin{aligned}
\mathsf{MSE}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_x[\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(x)\|_2^2] \\
&= \mathbb{E}_x\left\{\left\|\boldsymbol{\theta} - \mathbb{E}[\hat{\boldsymbol{\theta}}(x)] + \mathbb{E}[\hat{\boldsymbol{\theta}}(x)] - \hat{\boldsymbol{\theta}}(x)\right\|_2^2\right\} \\
&= \|\boldsymbol{\theta} - \mathbb{E}[\hat{\boldsymbol{\theta}}(x)]\|_2^2 + \mathbb{E}[\|\hat{\boldsymbol{\theta}}(x) - \mathbb{E}[\hat{\boldsymbol{\theta}}(x)]\|_2^2] \\
&\quad + 2(\boldsymbol{\theta} - \mathbb{E}[\hat{\boldsymbol{\theta}}(x)])^T \mathbb{E}[\hat{\boldsymbol{\theta}}(x) - \mathbb{E}[\hat{\boldsymbol{\theta}}(x)]] \\
&= \|\mathsf{bias}(\hat{\boldsymbol{\theta}})\|_2^2 + \mathsf{var}(\hat{\boldsymbol{\theta}})
\end{aligned}
$$

- Bia-Variance Decomposition
  The MSE is contributed by two parts:
  - Bias: $\|\mathsf{bias}(\hat{\boldsymbol{\theta}})\|_2^2$
  - Variance: $\mathsf{var}(\hat{\boldsymbol{\theta}})$

Example

$X_1, X_2, \ldots, X_n$ are i.i.d. random variables with pdf $\mathcal{N}(\mu, 1)$, where μ is an unknown parameter to estimate. Consider an estimator

$$\hat{\mu}_n = \hat{\mu}(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

What is the bias, variance of the estimator? What is the MSE?

Asymptotics

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with pdf $p(x|\theta), \theta \in \Theta$, and consider an estimator $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \ldots, X_n)$. How does $\hat{\theta}_n$ behave as $n \to \infty$?

Definition

$\hat{\theta}_n$ is asymptotically unbiased if $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] - \theta = 0$ for all $\theta \in \Theta$.

Definition

$\hat{\theta}_n$ is consistent (w.r.t chosen loss/risk) if $\lim_{n\to\infty} R(\theta, \hat{\theta}_n) = 0$ for all $\theta \in \Theta$.

Asymptotics

Example

$X_1, X_2, \ldots, X_n$ are i.i.d. random variables with pdf $\mathcal{N}(\mu, 1)$, where μ is an unknown parameter to estimate. Consider an estimator

$$\hat{\mu}_n = \hat{\mu}(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Consider ℓ_2 loss function $\ell(\mu, \hat{\mu}_n) = \|\mu - \hat{\mu}_n\|_2^2$, and the risk as $R(\mu, \hat{\mu}_n) = \mathbb{E}_x[\ell(\mu, \hat{\mu}_n)]$.

Maximum Likelihood Estimation

- The maximum Likelihood (ML) Estimate is given by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p(x|\theta)$$

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{p(x|\theta)} = \arg\min_{\theta \in \Theta} -\log p(x|\theta)$$

Example

Given a single observation of x generated according to $p(x|\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$. What is the MLE? Is it biased?

Solutions:

$$J(\theta) = -\log(p(x|\theta)) = \log\theta + \frac{x}{\theta}$$
$$\frac{dJ(\theta)}{d\theta} = \frac{1}{\theta} - \frac{x}{\theta^2} = \frac{1}{\theta}\left(1 - \frac{x}{\theta}\right)$$

- If $\theta < x$, then $\frac{dJ(\theta)}{d\theta} < 0$, that is, $J(\theta)$ decreases in θ
- If $\theta > x$, then $\frac{dJ(\theta)}{d\theta} > 0$, that is, $J(\theta)$ increases in θ
- Thus $J(\theta)$ is quasi-convex in θ, and achieves the minimum at $\frac{dJ(\theta)}{d\theta} = 0$

$$\hat{\theta}_{\mathsf{ML}} = x$$

.

Example

Given $p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T\Sigma^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})\}$, $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{\theta} \in \mathbb{R}^k$.
What is the MLE of $\boldsymbol{\theta}$?

Solutions:

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T\Sigma^{-1}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{x}^T\Sigma^{-1}\mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{H}^T\Sigma^{-1}\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\boldsymbol{\theta} \\ \frac{\partial J(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= -2\mathbf{H}^T\Sigma^{-1}\mathbf{x} + 2\mathbf{H}^T\Sigma^{-1}\mathbf{H}\boldsymbol{\theta} = 0 \end{aligned}$$

$$\hat{\boldsymbol{\theta}}_{\mathsf{ML}} = (\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\Sigma^{-1}\mathbf{x}$$

Example

Consider $x[n] = A + w[n], n = 0, 1, \ldots, N$, where $w[n]$ is WGN with variance σ^2. Find MLE for the vector parameter $\boldsymbol{\theta} = [A, \sigma^2]^T$. Is it unbiased?

Solution: p. 183, Example 7.12, Kay Volume 1

Example

Consider $x[n] = A + w[n], n = 0, 1, \ldots, N$, where $w[n]$ is WGN with variance σ^2. Show that the following estimator is an unbiased estiamte of the vector parameter $\boldsymbol{\theta} = [A, \sigma^2]^T$.

$$\hat{A} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{A})^2$$

Asymptotic Distribution of the MLE

Theorem

Let $x_1, x_2, \ldots, x_n$ *be i.i.d observations generated according to* $p(x|\boldsymbol{\theta}^*)$*, where* $\boldsymbol{\theta}^* \in \mathbb{R}^d$*. Let*

$$\hat{\boldsymbol{\theta}}_n := \arg\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

and $L(\boldsymbol{\theta}) := \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$*. Assume* $\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j}$ *and* $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}$ *exist for all* j, k*. Then,*

$$\hat{\boldsymbol{\theta}}_n \sim \mathcal{N}\left(\boldsymbol{\theta}^*, I^{-1}(\boldsymbol{\theta}^*)\right) \quad \textit{asymptotically}$$

where $I(\boldsymbol{\theta}^*)$ *is the Fisher-Information Matrix whose elements are given by*

$$[I(\boldsymbol{\theta}^*)]_{i,j} = -\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]$$

Asymptotic Distribution of the MLE

Example

Consider $x[n] = A + w[n], n = 0, 1, \ldots, N$, where $w[n]$ is WGN with variance σ^2. Find the asymptotics of the MLE estimate of $\boldsymbol{\theta} = [A, \sigma^2]^T$.

Solution: p. 183, Theorem 7.3, Kay Volume 1

Let $\theta_1 = A$ and $\theta_2 = \sigma^2$.

- $\log p(\mathbf{x}|\boldsymbol{\theta}) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i-1}^{N}(X_i - A)^2$

$$\frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_1} = \frac{1}{\sigma^2}\sum_{i-1}^{N}(X_i - A)$$

$$\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_1^2}\right] = \mathbb{E}\left[-\frac{N}{\sigma^2}\right] = -\frac{N}{\sigma^2}$$

$$\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2}\right] = \mathbb{E}\left[-\frac{1}{\sigma^4}\sum_{i=1}^{N}(X_i - A)\right] = 0$$

Solution:(Cont'd)

$$\frac{\partial \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i-1}^{N}(X_i - A)^2$$

$$\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2^2}\right] = \mathbb{E}\left[\frac{N}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}\sum_{i-1}^{N}(X_i - A)^2\right] = -\frac{N}{2\sigma^4}$$

$$\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1}\right] = \mathbb{E}\left[-\frac{1}{\sigma^4}\sum_{i-1}^{N}(X_i - A)\right] = 0$$

- Fisher Information matrix $I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$
- $I^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}$
- The exact coveriance matrix of $\hat{\boldsymbol{\theta}}$ is $\mathbf{C}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)\sigma^4}{N^2} \end{bmatrix} \sim I^{-1}(\boldsymbol{\theta})$

MLE for Transformed Parameters

In many instances, we wish to estimate a function of θ.

Example

Let $x_1, x_2, \ldots, x_n$ be be generated according to $x_i = A + W_i$, where W_i are WGN. Find the MLE of $\alpha = \exp(A)$.

Solution: Since $p(\mathbf{x}|A) \sim \mathcal{N}(A, \sigma^2)$, and α is a one-to-one transformation of A, we can equivalently parameterize the pdf as

$$p_T(\mathbf{x}|\alpha) \sim \mathcal{N}(\log \alpha, \sigma^2)$$

The MLE of α is found by maximizing $p_T(\mathbf{x}|\alpha)$.

Solution: (Cont'd)

$$p_T(\mathbf{x}|\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i - \log\alpha)^2\right)$$

$$\frac{\partial \log p_T(\mathbf{x}|\alpha)}{\partial \alpha} = \frac{1}{\sigma^2 \alpha}\sum(X_i - \log\alpha) = 0$$

$$\log\alpha = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X} = \hat{A}_{\mathsf{ML}}$$

$$\hat{\alpha}_{\mathsf{ML}} = \exp(\hat{A}_{\mathsf{ML}})$$

The MLE of the transformed parameter is found by substituting the MLE of the original parameter into the transformation.

Example

Now consider the transformation $\alpha = A^2$ for the previous example.

Since $A = +/-\sqrt{\alpha}$, the transformation is not one-to-one.
If $A = \sqrt{\alpha}$, $p_{t1}(x|\alpha) \sim \mathcal{N}(\sqrt{\alpha}, \sigma^2)$.
If $A = -\sqrt{\alpha}$, $p_{t1}(x|\alpha) \sim \mathcal{N}(-\sqrt{\alpha}, \sigma^2)$.
Then, the MLE of α is

$$\hat{\alpha} = \arg\max_{\alpha} \big(p_{t1}(\mathbf{x}|\alpha), p_{t2}(\mathbf{x}|\alpha)\big)$$

Invariance of the MLE

Theorem

The MLE of the parameter $\alpha = g(\theta)$, where the pdf $p(x|\theta)$ is parameterized by θ is given by

$$\hat{\tau} = g(\hat{\theta})$$

where $\hat{\theta}$ is the MLE of θ. If g is not a non-to-one function, then $\hat{\alpha}$ maximizes the modified likelihood function $p_t(x|\alpha)$ defined as

$$p_t(x|\alpha) = \max_{\theta:\alpha=g(\theta)} p(x|\theta)$$