ECE531 Lecture 10a: Maximum Likelihood Estimation

D. Richard Brown III

Worcester Polytechnic Institute

02-Apr-2009

Introduction

- So far, we have two techniques for finding "good" estimators when the parameter is non-random:
  1. Rao-Blackwell-Lehmann-Sheffe technique for finding MVU estimators.
  2. "Guess and check": use your intuition to guess at a good estimator and then compare variance to information inequality or CRLB
- Both techniques can fail or be difficult to use in some scenarios, e.g.
  - Can't find a complete sufficient statistic
  - Can't compute the conditional expectation in the RBLS theorem
  - Can't invert the Fisher information matrix
  - ...
- Today we will learn a third technique that can often help us find good estimators: the **maximum likelihood** criterion.

The Maximum Likelihood Criterion

- Suppose for a moment that our parameter Θ is random (like Bayesian estimation) and we have a prior on the unknown parameter $\pi_\Theta(\theta)$. The maximum a posteriori (MAP) estimator can be written as

$$\hat{\theta}_{\mathsf{map}}(y) = \arg\max_{\theta \in \Lambda} p_\Theta(\theta \,|\, y) \stackrel{\text{Bayes}}{=} \arg\max_{\theta \in \Lambda} p_Y(y \,|\, \theta)\pi_\Theta(\theta)$$

- If you were unsure about this prior, you might assume a "least favorable prior". Intuitively, what sort of prior would give the least information about the parameter?
- We could assume the prior is **uniformly distributed** over Λ, i.e. $\pi_\Theta(\theta) \equiv \pi > 0$ for all $\theta \in \Lambda$. Then

$$\hat{\theta}_{\mathsf{map}}(y) = \arg\max_{\theta \in \Lambda} p_Y(y\,;\,\theta)\pi = \arg\max_{\theta \in \Lambda} p_Y(y\,;\,\theta) = \hat{\theta}_{\mathsf{ml}}(y)$$

  finds the value of θ that **makes the observation $Y = y$ most likely**. $\hat{\theta}_{\mathsf{ml}}(y)$ is called the **maximum likelihood** (ML) estimator.
- Problem 1: If Λ is not a bounded set, e.g. $\Lambda = \mathbb{R}$, then $\pi = 0$. A uniform distribution on $\mathbb{R}$ (or any unbounded Λ) doesn't really make sense.

The Maximum Likelihood Criterion

- Even though our development of the ML estimator is questionable, the criterion of finding the value of θ that makes the observation $Y = y$ most likely (assuming all parameter values are equally likely) is still interesting.
- Note that, since $\ln$ is strictly monotonically increasing,

$$\hat{\theta}_{\mathsf{ml}}(y) = \arg\max_{\theta \in \Lambda} p_Y(y\,;\,\theta) = \arg\max_{\theta \in \Lambda} \ln p_Y(y\,;\,\theta)$$

- Assuming that $\ln p_Y(y\,;\,\theta)$ is differentiable, we can state a **necessary** (but not sufficient) condition for the ML estimator:

$$\boxed{\left.\frac{\partial}{\partial\theta} \ln p_Y(y\,;\,\theta)\right|_{\theta=\hat{\theta}_{\mathsf{ml}}(y)} = 0}$$

  where the partial derivative is taken to mean the gradient ∇_θ for multiparameter estimation problems.
- This expression is known as the **likelihood equation**.

The Likelihood Equation's Relationship with the CRLB

- Recall from last week's lecture that we can attain the CRLB if and only if $p_Y(y\,;\,\theta)$ is of the form

$$p_Y(y\,;\,\theta) = h(y)\exp\left\{\int_a^\theta I(t)\left[\hat{\theta}(y)-\theta\right]\,dt\right\}$$
$$\Leftrightarrow \quad \ln p_Y(y\,;\,\theta) = \ln h(y) + \int_a^\theta I(t)\left[\hat{\theta}(y)-\theta\right]\,dt$$

  for all $y \in \mathcal{Y}$.
- When this condition holds, the likelihood equation becomes

$$\frac{\partial}{\partial\theta}\ln p_Y(y\,;\,\theta)\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}(y)} = I(\theta)\left[\hat{\theta}(y)-\theta\right]_{\theta=\hat{\theta}_{\mathsf{ml}}(y)} = 0$$

  which, as long as $I(\theta)>0$, has the unique solution $\hat{\theta}_{\mathsf{ml}}(y)=\hat{\theta}(y)$.
- What does this mean? If $\hat{\theta}(y)$ attains the CRLB, it must be a solution to the likelihood equation. The converse is not always true, however.

Some Initial Properties of Maximum Likelihood Estimators

- If $\hat{\theta}(y)$ attains the CRLB, it must be a solution to the likelihood equation.
  - In this case, $\hat{\theta}_{\mathsf{ml}}(y) = \hat{\theta}_{\mathsf{mvu}}(y)$.
- Solutions to the likelihood equation may not achieve the CRLB.
  - In this case, it may be possible to find other unbiased estimators with lower variance than the ML estimator.

Example: Estimating a Constant in White Gaussian Noise

Suppose we have random observations given by

$$Y_k \quad = \quad \theta + W_k \qquad k = 0, \dots, n-1$$

where $W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. The unknown parameter θ is non-random and can take on any value on the real line (we have no prior pdf).

Let's set up the likelihood equation $\frac{\partial}{\partial\theta} \ln p_Y(y\,;\,\theta)\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}(y)} = 0$...

We can write

$$\frac{\partial}{\partial\theta} \ln \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-\sum_{k=0}^{n-1}(y_k-\theta)^2}{2\sigma^2}\right\}\Bigg|_{\theta=\hat{\theta}_{\mathsf{ml}}} = \frac{1}{\sigma^2}\sum_{k=0}^{n-1}(y_k - \hat{\theta}_{\mathsf{ml}}) = 0$$

hence

$$\hat{\theta}_{\mathsf{ml}}(y) \quad = \quad \frac{1}{n}\sum_{k=0}^{n-1} y_k = \bar{y}.$$

This solution is unique. You can also easily confirm that it maximizes $\ln p_Y(y\,;\,\theta)$.

Example: Est. the Parameter of an Exponential Distrib.

Suppose we have random observations given by $Y_k \overset{\text{i.i.d.}}{\sim} \theta e^{-\theta y_k}$ for $k = 0, \dots, n-1$ and $y_k \geq 0$. The unknown parameter $\theta > 0$ and we have no prior pdf.

Let's set up the likelihood equation $\frac{\partial}{\partial\theta} \ln p_Y(y\,;\,\theta)\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}(y)} = 0$...

Since $p_Y(y\,;\,\theta) = \prod_k p_{Y_k}(y_k\,;\,\theta) = \theta^n \exp\left\{-\theta \sum_k y_k\right\}$, we can write

$$\begin{aligned} \frac{\partial}{\partial\theta} \ln\left[\theta^n \exp\left\{-n\theta\bar{y}\right\}\right]\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}} &= \frac{\partial}{\partial\theta}\left(n\ln\theta - n\theta\bar{y}\right)\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}} \\ &= \frac{n}{\hat{\theta}_{\mathsf{ml}}(y)} - n\bar{y} = 0 \end{aligned}$$

which has the unique solution $\hat{\theta}_{\mathsf{ml}}(y) = 1/\bar{y}$. You can easily confirm that this solution is a maximum of $p_Y(y\,;\,\theta)$ since $\frac{\partial^2}{\partial\theta^2} \ln p_Y(y\,;\,\theta) = \frac{-n}{\theta^2} < 0$.

Example: Est. the Parameter of an Exponential Distrib.

Assuming $n \ge 2$, the mean of the ML estimator can be computed as

$$\mathrm{E}_\theta\left\{\hat{\theta}_{\mathsf{ml}}(Y)\right\} \quad = \quad \frac{n}{n-1}\theta$$

Note that $\hat{\theta}_{\mathsf{ml}}(y)$ is biased for finite n but it is **asymptotically unbiased** in the sense that $\lim_{n\to\infty} \mathrm{E}_\theta\left\{\hat{\theta}_{\mathsf{ml}}(Y)\right\} = \theta$. Assuming $n \ge 3$, the variance of the ML estimator can be computed as

$$\mathrm{var}_\theta\left\{\hat{\theta}_{\mathsf{ml}}(Y)\right\} = \frac{n^2\theta^2}{(n-1)^2(n-2)} \quad > \quad \frac{\theta^2}{n} = I^{-1}(\theta)$$

where $I(\theta)$ is the Fisher information for this estimation problem.

- Since $\frac{\partial}{\partial\theta}\ln p_Y(y\,;\,\theta)$ is not of the form $k(\theta)\left[\hat{\theta}_{\mathsf{ml}}(y) - f(\theta)\right]$, we know that the information inequality can not be attained here.
- Nevertheless, $\hat{\theta}_{\mathsf{ml}}(y)$ is **asymptotically efficient** in the sense that $\lim_{n\to\infty} \mathrm{var}_\theta\left\{\hat{\theta}_{\mathsf{ml}}(Y)\right\} = I^{-1}(\theta)$.

Example: Est. the Parameter of an Exponential Distrib.

It can be shown using the RBLS theorem and our theorem about complete sufficient statistics for exponential density families that

$$\hat{\theta}_{\text{mvu}}(y) \quad = \quad \left(\frac{1}{n-1} \sum_{k=0}^{n-1} y_k \right)^{-1} = \frac{n-1}{n} \hat{\theta}_{\text{ml}}(y).$$

Assuming $n \geq 3$, the variance of the MVU estimator is then

$$\text{var}_\theta \left\{ \hat{\theta}_{\text{mvu}}(Y) \right\} = \frac{\theta^2}{n-2} \quad > \quad \frac{\theta^2}{n} = I^{-1}(\theta).$$

Which is better, $\hat{\theta}_{\text{ml}}(y)$ or $\hat{\theta}_{\text{mvu}}(y)$?

Which is Better, MVU or ML?

To answer this question, you have to return to what got us here: the **squared error** cost function.

$$
\begin{aligned}
\mathrm{E}_\theta\left\{(\hat{\theta}_{\mathsf{ml}}(Y)-\theta)^2\right\} &= \mathrm{var}_\theta\left\{\hat{\theta}_{\mathsf{ml}}(Y)\right\}+\left(\frac{\theta}{n-1}\right)^2=\frac{n+2}{n-1}\cdot\frac{\theta^2}{n-2}\\
&> \frac{\theta^2}{n-2}=\mathrm{var}_\theta\left\{\hat{\theta}_{\mathsf{mvu}}(Y)\right\}.
\end{aligned}
$$

The MVU estimator is preferable to the ML estimator for finite n. Asymptotically, however, their squared error performance is equivalent.

Example: Estimating a The Mean and Variance of WGN

Suppose we have random observations given by $Y_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $k = 0, \ldots, n-1$. The unknown vector parameter $\theta = [\mu, \sigma^2]$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The joint density is given as

$$p_Y(y\,;\,\theta) \quad = \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{-\sum_{k=0}^{n-1}(y_k - \mu)^2}{2\sigma^2} \right\}$$

How should we approach this joint maximization problem? In general, we would compute the gradient $\nabla_\theta \ln p_Y(y\,;\,\theta)\Big|_{\theta=\hat{\theta}_{\mathsf{ml}}(y)}$, set the result equal to zero, and then try to solve the simultaneous equations (also using the Hessian to check that we did indeed find a maximum)...

Or we could recognize that finding the value of μ that maximizes $p_Y(y\,;\,\theta)$ does not depend on σ^2...

Example: Estimating a The Mean and Variance of WGN

We already know $\hat{\mu}_{\sf ml}(y)$ for estimating μ:

$$\hat{\mu}_{\sf ml}(y) \quad = \quad \bar{y} \qquad \text{(same as MVU estimator).}$$

So now we just need to solve

$$\hat{\sigma}^2_{\sf ml} \quad = \quad \arg\max_{a>0} \left(\ln \left\{ \frac{1}{(2\pi a)^{n/2}} \exp \left\{ \frac{-\sum_{k=0}^{n-1}(y_k - \bar{y})^2}{2a} \right\} \right\} \right)$$

Skipping the details (standard calculus, see example IV.D.2 in your textbook), it can be shown that

$$\hat{\sigma}^2_{\sf ml} = \frac{1}{n} \sum_{k=0}^{n-1} (y_k - \bar{y})^2$$

Example: Estimating a The Mean and Variance of WGN

Is $\hat{\sigma}^2_{\mathsf{ml}}$ biased in this case?

$$\begin{aligned}
\mathrm{E}_\theta\left\{\hat{\sigma}^2_{\mathsf{ml}}(Y)\right\} &= \frac{1}{n}\sum_{k=0}^{n-1}\mathrm{E}_\theta[(Y_k-\mu)^2 - 2(Y_k-\mu)(\bar{Y}-\mu) + (\bar{Y}-\mu)^2] \\
&= \frac{1}{n}\sum_{k=0}^{n-1}\sigma^2 - \frac{2\sigma^2}{n} + \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Yes, $\hat{\sigma}^2_{\mathsf{ml}}$ is biased but it is asymptotically unbiased.

The steps in the previous analysis can be followed to show that $\hat{\sigma}^2_{\mathsf{ml}}$ is unbiased if μ is known. The unknown mean, even though we have an unbiased efficient estimator of it, causes $\hat{\sigma}^2_{\mathsf{ml}}(y)$ to be biased here.

Example: Estimating a The Mean and Variance of WGN

It can also be shown that

$$\mathrm{var}\left\{\hat{\sigma}^2_{\mathsf{ml}}(Y)\right\} = \frac{2(n-1)\sigma^4}{n^2} \quad < \quad \frac{2\sigma^4}{n-1} = \mathrm{var}\left\{\hat{\sigma}^2_{\mathsf{mvu}}(Y)\right\}$$

Which is better, $\hat{\sigma}^2_{\mathsf{ml}}(y)$ or $\hat{\sigma}^2_{\mathsf{mvu}}(y)$? To answer this, let's compute the mean squared error (MSE) of the ML estimator for σ^2:

$$\begin{aligned}
\mathrm{E}_\theta\left\{(\hat{\sigma}^2_{\mathsf{ml}}(Y)-\sigma^2)^2\right\} &= \mathrm{var}_\theta\left\{\hat{\sigma}^2_{\mathsf{ml}}(Y)\right\} + \left(\mathrm{E}_\theta\left\{\hat{\sigma}^2_{\mathsf{ml}}(Y)\right\}-\sigma^2\right)^2 \\
&= \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2-\sigma^2\right)^2 \\
&= \frac{(2n-1)\sigma^4}{n^2} \\
&< \frac{2\sigma^4}{n-1} = \mathrm{E}_\theta\left\{(\hat{\sigma}^2_{\mathsf{mvu}}(Y)-\sigma^2)^2\right\}
\end{aligned}$$

Hence the ML estimator has uniformly lower mean squared error (MSE) performance than the MVU estimator. The increase in MSE due to the bias is more than offset by the decreased variance of the ML estimator.

More Properties of ML Estimators

From our examples, we can say that

- Maximum likelihood estimators may be biased.
- A biased ML estimator may, in some cases, outperform a MVU estimator in terms of overall mean squared error.
- It seems that ML estimators are often asymptotically unbiased:

$$\lim_{n\to\infty} \mathrm{E}_\theta \left\{ \hat{\theta}_{\mathsf{ml}}(Y) \right\} = \theta$$

  Is this always true?
- It seems that ML estimators are often asymptotically efficient:

$$\lim_{n\to\infty} \mathrm{var}_\theta \left\{ \hat{\theta}_{\mathsf{ml}}(Y) \right\} = I^{-1}(\theta)$$

  Is this always true?

Consistency of ML Estimators For i.i.d. Observations

Suppose that we have i.i.d. observations with marginal distribution $Y_k \overset{\text{i.i.d.}}{\sim} p_Z(z\,;\theta)$ and define

$$\begin{aligned}\psi(z\,;\theta') &:= \left.\frac{\partial}{\partial\theta}\ln p_Z(z\,;\theta)\right|_{\theta=\theta'} \\ J(\theta\,;\theta') &:= \mathrm{E}_\theta\left\{\psi(Y_k\,;\theta')\right\} = \int_{\mathcal{Z}} \psi(z\,;\theta')p_Z(z\,;\theta)\,dz\end{aligned}$$

where $\mathcal{Z}$ is the support of the pdf of a single observation $p_Z(z\,;\theta)$.

Theorem

If all of the following (sufficient but not necessary) conditions hold

1. $J(\theta\,;\theta')$ *is a continuous function of* θ' *and has a unique root* $\theta' = \theta$ *at which point* J *changes sign,*
2. $\psi(Y_k\,;\theta')$ *is a continuous function of* θ' *(almost surely), and*
3. *For each* n, $\frac{1}{n}\sum_{k=0}^{n-1}\psi(Y_k\,;\theta')$ *has a unique root* $\hat{\theta}_n$ *(almost surely),*

then $\hat{\theta}_n \xrightarrow{\text{i.p.}} \theta$, *where* $\xrightarrow{\text{i.p.}}$ *means convergence in probability.*

Consistency of ML Estimators For i.i.d. Observations

Convergence in probability means that

$$\lim_{n\to\infty} \mathrm{Prob}_\theta \left[\left| \hat{\theta}_n(Y) - \theta \right| > \epsilon \right] = 0 \text{ for all } \epsilon > 0$$

Example: Estimating a constant in white Gaussian noise: $Y_k = \theta + W_k$ for $k = 0, \dots, n-1$ with $W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. For finite n, we know the estimator $\hat{\theta}_n(y) = \bar{y} \sim \mathcal{N}(\theta, \sigma^2/n)$. Hence

$$\mathrm{Prob}_\theta \left[\left| \hat{\theta}_n(Y) - \theta \right| > \epsilon \right] = 2Q\left(\frac{\sqrt{n}\epsilon}{\sigma} \right)$$

and

$$\lim_{n\to\infty} 2Q\left(\frac{\sqrt{n}\epsilon}{\sigma} \right) = 0$$

for any $\epsilon > 0$. So this estimator (ML=MVU here) is consistent.

Consistency of ML Estimators: Intuition

Setting likelihood equation equal to zero for i.i.d. observations:

$$\frac{\partial}{\partial\theta} \ln p_Y(y\,;\,\theta)\Big|_{\theta=\theta'} \;=\; \sum_{k=0}^{n-1} \psi(y_k\,;\,\theta') = 0 \tag{1}$$

The **weak law of large numbers** tells us that

$$\frac{1}{n}\sum_{k=0}^{n-1} \psi(Y_k\,;\,\theta') \xrightarrow{\text{i.p.}} \mathrm{E}_\theta\left\{\psi(Y_1\,;\,\theta')\right\} = J(\theta\,;\,\theta')$$

Hence, solving (1) in the limit as $n \to \infty$ is equivalent to finding θ' such that $J(\theta\,;\,\theta') = 0$. Under our usual regularity conditions, it is easy to show that $J(\theta\,;\,\theta')$ will always have a root at $\theta' = \theta$.

$$\begin{aligned} J(\theta\,;\,\theta')\Big|_{\theta'=\theta} &= \int_{\mathcal{Z}} \frac{\partial}{\partial\theta} \ln p_Z(z\,;\,\theta) p_Z(z\,;\,\theta)\,dz \\ &= \int_{\mathcal{Z}} \frac{\frac{\partial}{\partial\theta} p_Z(z\,;\,\theta)}{p_Z(z\,;\,\theta)} p_Z(z\,;\,\theta)\,dz = \frac{\partial}{\partial\theta}\int_{\mathcal{Z}} p_Z(z\,;\,\theta)\,dz = 0 \end{aligned}$$

Asymptotic Unbiasedness of ML Estimators

An estimator is asymptotically unbiased if

$$\lim_{n\to\infty} \mathrm{E}_\theta\left\{\hat{\theta}_{\mathsf{ml},n}(Y)\right\} = \theta$$

Asymptotically unbiasedness requires **convergence in mean**. We've already seen several examples of ML estimators that are asymptotically unbiased.

Consistency implies that

$$\mathrm{E}_\theta\left\{\lim_{n\to\infty} \hat{\theta}_{\mathsf{ml},n}(Y)\right\} = \theta$$

This is **convergence in probability**. Does consistency imply asymptotic unbiasedness? Only when the limit and expectation can be exchanged.

- In most cases, this exchange is valid (see email sent last Friday).
- If you want to know more precisely the conditions under which this exchange is valid, you need to learn about "dominated convergence" (a course in Analysis will cover this).

Asymptotic Normality of ML Estimators

Main idea: For i.i.d. observations each distributed as $p_Z(z\,;\,\theta)$ and under regularity conditions similar to those you've seen before,

$$\boxed{\sqrt{n}(\hat{\theta}_{\mathsf{ml},n}(Y)-\theta)\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{i(\theta)}\right)}$$

where $\xrightarrow{d}$ means convergence in distribution and

$$i(\theta) \quad := \quad \mathrm{E}_\theta\left\{\left[\frac{\partial}{\partial\theta}\ln p_Z(Z\,;\,\theta)\right]^2\right\}$$

is the Fisher information of a single observation Y_k about the parameter θ.

See Proposition IV.D.2 in your textbook for the details.

Asymptotic Normality of ML Estimators

- The asymptotic normality of ML estimators is very important.
- For similar reasons that convergence in probability is not sufficient to imply asymptotic unbiasedness, convergence in distribution is also not sufficient to imply that

$$\mathrm{E}_\theta[\sqrt{n}(\hat{\theta}_n(Y) - \theta)] \to 0 \text{ (asymptotic unbiasedness)}$$

or

$$\mathrm{var}_\theta[\sqrt{n}(\hat{\theta}_n(Y) - \theta)] \to \frac{1}{i(\theta)} \text{ (asymptotic efficiency).}$$

- Nevertheless, as our examples showed, ML estimators are often both asymptotically unbiased and asymptotically efficient.
- When an estimator is asymptotically unbiased and asymptotically efficient, it is asymptotically MVU.

Conclusions

- Even though we started out with some questionable assumptions to specify the ML estimator, we found that that ML estimators can have nice properties:
  - If an estimator achieves the CRLB, then it must be a solution to the likelihood equation. Note that the converse is not always true.
  - For i.i.d. observations, ML estimators are guaranteed to be **consistent** (within regularity).
  - For i.i.d. observations, ML estimators are **asymptotically Gaussian** (within regularity).
  - For i.i.d. observations, ML estimators are often **asymptotically unbiased** and **asymptotically efficient**.
- Unlike MVU estimators, ML estimators may be biased.
- It is customary in many problems to assume sufficiently large n such that the asymptotic properties hold, at least approximately.
- Extensions to vector parameters can be found in the Poor textbook.