LECTURE 9A: MAXIMUM LIKELIHOOD ESTIMATION

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
OCTOBER 23, 2024

1. Cramer-Rao lower bound

An estimator is a *best unbiased estimator* if it achieves the lowest variance among all possible unbiased estimators.

The Cramer-Rao lower bound gives a lower bound on the variance of unbiased estimators. If an unbiased estimator achieves the Cramer-Rao lower bound, then it is a best (most efficient) unbiased estimator.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the density $f(x|\theta)$. Define the Fisher's Information number as:

$$\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right] \tag{1}$$

Where the expectation is taken with respect to $X \sim f(x|\theta)$. That is,

$$\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right] = n\int\left(\frac{\partial \log f(x|\theta)}{\partial \theta}\right)^2 f(x|\theta)\,dx$$

The Fisher's information number varies with the actual parameter. The higher the Fisher's information number, the lower the minimum achievable variance. Intuitively, large magnitudes of $\frac{\partial \log f(x|\theta)}{\partial \theta}$ means that the log-likelihood is very sensitive to small changes in θ, which accordingly contains more information about θ.

Theorem 1 (Cramer-Rao Inequality (i.i.d case))**.** *Let $X_1, \ldots, X_n$ be a random sample i.i.d from the pdf $f(x|\theta)$, and let $T(\boldsymbol{X})$ be an estimator for θ such that certain regularity conditions on the density are satisfied. These regularity conditions are*

$\text{Var}(T(\boldsymbol{X})) < \infty$ *and*

$$\frac{d}{d\theta}\,\mathbb{E}[T(\boldsymbol{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} T(\boldsymbol{x}) f(\boldsymbol{x}|\theta) d\boldsymbol{x}$$

then the Cramer-Rao Inequality says that,

$$\text{Var}(T(\boldsymbol{X})) \geq \frac{(\frac{\partial}{\partial\theta}\,\mathbb{E}[T(\boldsymbol{X})])^2}{\mathcal{I}(\theta)} \tag{2}$$

$\mathcal{X}$ is the support of the pdf $f(x|\theta)$. This result holds for both discrete and continuous random variables. These regularity conditions are satisfied usually with the key exception of Uniform distribution.[1] Further, if $T(\boldsymbol{X})$ is an unbiased estimator for θ, then $\mathbb{E}[T(\boldsymbol{X})] = \theta$, and so that $\frac{\partial}{\partial\theta}\,\mathbb{E}[T(\boldsymbol{X})] = 1$. Among unbiased estimators, the Cramer-Rao lower bound becomes:

$$\text{Var}(T(\boldsymbol{X})) \geq \frac{1}{\mathcal{I}(\theta)} \tag{3}$$

The Fisher's Information number (and the Cramer-Rao inequality) is defined more generally as follows. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a (possibly non-iid) sample from the joint density $f(x_1, \ldots, x_n|\theta)$. Then,

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(\boldsymbol{X}|\theta)}{\partial\theta}\right)^2\right] \tag{4}$$

Where the expectation is taken with respect to the joint density $f(x_1, \ldots, x_n|\theta)$.

Theorem 2 (Cramer-Rao Inequality)**.** *Let $X_1, \ldots, X_n$ be a sample with pdf $f(x_1, \ldots, x_n|\theta)$, and let $T(\boldsymbol{X})$ be an estimator for θ such that* $\text{Var}(T(\boldsymbol{X})) < \infty$ *and*

$$\frac{d}{d\theta}\,\mathbb{E}[T(\boldsymbol{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} T(\boldsymbol{x}) f(\boldsymbol{x}|\theta) d\boldsymbol{x}$$

[1] The density is sufficiently smooth such that we can interchange integration and differentiation. See Leibniz integral rule, which requires that both $f(\boldsymbol{x}|\theta)$ and $\frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta}$ to be continuous in $\boldsymbol{x}$ and θ. This is guaranteed when either the function $f(x|\theta)$ has bounded support in x, and the bounds do not depend on θ, or the function $f(x|\theta)$ has infinite support and it is continuously differentiable. This regularity condition fails when the parameter space depends on the parameter, such as in the Uniform distribution when the support depends on the parameter.

then,

$$\text{Var}(T(\boldsymbol{X})) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}[T(\boldsymbol{X})])^2}{\mathcal{I}(\theta)} \tag{5}$$

The derivative of the log density $s(\theta|\boldsymbol{x}) = \frac{\partial \log f(\boldsymbol{x}|\theta)}{\partial\theta}$ is also called the *Score* function. Under some regularity conditions for $f(\boldsymbol{x}|\theta)$, we have $\mathbb{E}[s(\theta|\boldsymbol{X})] = 0$.

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \log f(\boldsymbol{X}|\theta)}{\partial\theta}\right] &= \int \frac{\partial \log f(\boldsymbol{x}|\theta)}{\partial\theta} f(\boldsymbol{x}|\theta) d\boldsymbol{x} \\
&= \int \frac{1}{f(\boldsymbol{x}|\theta)} \frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta} f(\boldsymbol{x}|\theta) d\boldsymbol{x} \\
&= \frac{\partial}{\partial\theta} \int f(\boldsymbol{x}|\theta) d\boldsymbol{x} \\
&= 0
\end{aligned}$$

Therefore, the Fisher Information is the variance of the score function, i.e. $\mathbb{E}\left[\left(\frac{\partial \log f(\boldsymbol{X}|\theta)}{\partial\theta}\right)^2\right] = \text{Var}\left(\frac{\partial \log f(\boldsymbol{X}|\theta)}{\partial\theta}\right)$. Fisher Information can be thought of as a measure of the curvature or "sharpness" of the likelihood function. When the likelihood function is very peaked (sharp) with respect to θ, then the Fisher Information is large, meaning that the observation carries a lot of information about θ. Conversely, if the likelihood function is flat, the Fisher Information is small, indicating that the observation carries little information about θ.

Finally, when the log density is twice differentiable with respect to θ, and under certain regularity conditions we have another equivalent formula to calculate Fisher information:

$$\mathcal{I}(\theta) = -n\mathbb{E}\left[\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}\right] = -n\int \frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} f(x|\theta)\, dx$$

1.1. Example

Consider the Poisson distribution $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$. Because $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$, both $\bar{X}$ and S^2 are unbiased estimators of the rate parameter λ. Which estimator should we use?

With the Poisson distribution, the Fisher's information number is:

$$
\begin{aligned}
\mathcal{I}(\lambda) &= n \sum_{x=0}^{\infty} \left(\frac{\partial \log f(x|\lambda)}{\partial \lambda} \right)^2 \frac{e^{-\lambda}\lambda^x}{x!} \\
&= n \sum_{x=0}^{\infty} \left(\frac{x}{\lambda} - 1 \right)^2 \frac{e^{-\lambda}\lambda^x}{x!} \\
&= n \sum_{x=0}^{\infty} \left(\frac{x^2}{\lambda^2} - \frac{2x}{\lambda} + 1 \right) \frac{e^{-\lambda}\lambda^x}{x!} \\
&= n \sum_{x=0}^{\infty} \left(\frac{x^2}{\lambda^2} - \frac{2x}{\lambda} \right) \frac{e^{-\lambda}\lambda^x}{x!} + n \\
&= n \sum_{x=0}^{\infty} \left(\frac{x^2}{\lambda^2} \right) \frac{e^{-\lambda}\lambda^x}{x!} - 2n + n \\
&= \frac{n}{\lambda^2}(\lambda + \lambda^2) - n \\
&= \frac{n}{\lambda}
\end{aligned}
$$

Recall that for the Poisson distribution with unknown parameter, $\bar{X}$ is an unbiased estimator of λ, and moreover $\text{Var}(\bar{X}) = \frac{\lambda}{n}$. Therefore, the estimator $\bar{X}$ for λ in the Poisson case achieves the Cramer-Rao lower bound, and hence, it is the best unbiased estimator of λ.

2. Properties of MLE

In finite-samples (for small n), the MLE is often dominated by other estimators in terms of unbiasedness or mean-squared error. However, MLE is widely used, as under certain regularity conditions, the MLE possesses many desirable properties. Notably, MLE is consistent, asymptotically Normal and it is efficient. Among all consistent and asymptotically normal estimators, the MLE achieves the smallest possible asymptotic variance, which is the inverse of the Fisher Information.

2.1. Consistency and identification

MLE is a consistent estimator. That is, $\hat{\theta}$ converges in probability to θ, the true parameter value, as $n \to \infty$. Therefore, whatever bias that MLE suffers, it will disappear when enough sample size is collected.

There are several regularity conditions that are sufficient for MLE to be consistent, efficient and asymptotically Normal. These conditions ensure that the likelihood

function behaves "nicely" in a neighborhood of the true parameter value. For example, (1) the likelihood function is sufficiently smooth and differentiable in the parameter, (2) the parameter space is compact (which essentially means that MLE exists), (3) the true parameter lies in the interior of the parameter space and not on its boundary, etc. In practice, the main substantive conditions we have to worry about is the *identification* condition, and that the density is correctly specified.

The parameters must be identified in the following sense. If $\theta \neq \theta'$, then $L(\theta|x) \neq L(\theta'|x)$. If this condition does not hold, then there are two parameter values that generate the same likelihood. We would not be able to distinguish between these two parameters even with an infinite amount of data, these parameters would have been observationally equivalent.

As an example, consider estimating the parameters α, μ, σ given the model $\mathcal{N}(\alpha\mu, \sigma^2)$. Then, (α, μ, σ) and $(k\alpha, \frac{\mu}{k}, \sigma)$ are observationally equivalent and have the same likelihood. If $(\hat{\alpha}, \hat{\mu}, \hat{\sigma}) = \operatorname{argmax} L(\alpha, \mu, \sigma)$, then $(k\hat{\alpha}, \hat{\mu}/k, \hat{\sigma}) = \operatorname{argmax} L(\alpha, \mu, \sigma)$ for any k. The parameters are not identified, MLE will not converge to the true parameter values. Is the model $\mathcal{N}(\mu - \alpha, \sigma^2/\alpha)$ identified? What about mixtures of Normals?

2.2. Asymptotic normality

MLE is asymptotically Normal. That is, $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$ converges in distribution to a Normal distribution with mean zero (since it is consistent).

Moreover, MLE is also the most efficient estimator when $n \to \infty$. It achieves the Cramer-Rao lower bound when n is large.

Specifically, suppose that $\hat{\theta}_{MLE}$ is an MLE of θ given the data $x_1, \ldots, x_n$ realizes i.i.d from $f(x|\theta)$.[2] We have that $\hat{\theta}_{MLE}$ is asymptotically Normal:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \xrightarrow{d} \mathcal{N}(0, n\mathcal{I}(\theta)^{-1}) \quad \text{as } n \to \infty$$

Where $\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]$ is the Fisher's information number. The variance of MLE is therefore approximated by $\mathcal{I}(\theta)^{-1}$. Now θ is unknown, but by Continuous Mapping Theorem, we can plug in a consistent estimator of θ, i.e. we can compute the asymptotic variance of MLE as $\mathcal{I}(\hat{\theta})^{-1}$, where $\hat{\theta}$ is the MLE of θ.[3]

[2] Assume that θ is a scalar, but the result generalizes to a vector of parameters.

[3] $\hat{\theta} \to_p \theta$ implies that $\mathcal{I}(\hat{\theta})^{-1} \to_p \mathcal{I}(\theta)^{-1}$

In addition, $\mathcal{I}(\theta)^{-1}$ often has no closed-form, so we estimate $\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]$ using its sample moment: $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\right)^2$.

2.3. Example

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, and consider the maximum likelihood of σ^2. We know that the MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, which has variance $\frac{2(n-1)\sigma^4}{n^2}$. If we calculate the Fisher's information number $\mathcal{I}(\theta) = n\,\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]$, we will find that $\mathcal{I}(\theta)^{-1} = \frac{2\sigma^4}{n}$. Therefore for asymptotically large n, the MLE of σ^2 achieves the Cramer-Rao lower bound variance.

2.4. Multiple parameters

When there are multiple parameters in the MLE,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, n\mathcal{I}^{-1}) \quad \text{as } n \to \infty$$

Where $\mathcal{I}$ is the Fisher's Information Matrix (for i.i.d. data-generating process):

$$\big[\mathcal{I}(\boldsymbol{\theta})\big]_{i,j} = n\,\mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i}\log f(X|\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \theta_j}\log f(X|\boldsymbol{\theta})\right)\right]$$

The vector $\boldsymbol{\nabla} \log f(X|\boldsymbol{\theta})$ is the gradient or the score function. The information matrix is then expectation of the outer product of the score vector, $\mathbb{E}[(\boldsymbol{\nabla} \log f(X|\boldsymbol{\theta}))(\boldsymbol{\nabla} \log f(X|\boldsymbol{\theta}))^T]$

We use $\mathcal{I}^{-1}$ as an (asymptotic) approximation of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{MLE}$.

In the multivariate case, the Cramer-Rao inequality becomes the following: let $T(\boldsymbol{X})$ be an unbiased estimator of $\boldsymbol{\theta}$, and let V be the variance-covariance matrix of $T(\boldsymbol{X})$, then $V - \mathcal{I}(\boldsymbol{\theta})^{-1}$ is positive definite.