LECTURE 9: BAYESIAN INFERENCE

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 4, 2020

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 unbiased estimator.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the density $f(x|\theta)$, and let $T(\boldsymbol{X})$ be an estimator for θ. 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. It measures the amount of information that a random variable X carries about an unknown parameter θ in the model $f(x|\theta)$. The higher the Fisher's information number, the lower the minimum achievable variance. Under some regularity conditions on f, the Cramer-Rao inequality says that[1], for any estimator $T(\boldsymbol{X})$:

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

[1]There are some cases where the Cramer-Rao inequality does not apply, for example, when the parameter space depends on the parameter

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:

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

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:

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

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.

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,

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

Note that $\log f(\boldsymbol{X}|\theta)$ is just the log-likelihood function $\mathcal{L}(\theta|x_1, \ldots, x_n)$ given realization $x_1, \ldots, x_n$.

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 the MLE has many attractive features when n is large.

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.

Although there are many technical conditions required in order to prove the consistency of MLE, many of these conditions cannot be falsified (such as the continuity of the likelihood function). In practice, the main substantive condition we have to worry about is the *identification* condition.

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, $(1, \mu, 1)$ and $(2, \frac{\mu}{2}, 1)$ are observationally equivalent. The parameter α is 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:

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

Where $\mathcal{I}(\theta)$ is the Fisher's information number. The variance of MLE is therefore approximated by $\frac{1}{n}\mathcal{I}(\theta)^{-1}$.

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 $\frac{2\sigma^2}{n-1}$. 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.

We rely on asymptotic approximations and take $\frac{1}{n}\mathcal{I}(\theta)^{-1}$ to be the variance of the MLE. However, often $\frac{1}{n}\mathcal{I}(\theta)^{-1}$ 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: $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \log f(x_i|\theta)}{\partial \theta}\right)^2$, or we estimate $\mathbb{E}\left[\left(\frac{\partial \log f(X_1,\ldots,X_n|\theta)}{\partial \theta}\right)^2\right]$ using a single draw: $\left(\frac{\partial \log f(x_1,\ldots,x_n|\theta)}{\partial \theta}\right)^2$

When there are multiple parameters in the MLE,

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

Where $\mathcal{I}$ is the Fisher's Information Matrix:

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

Therefore $\frac{1}{n}\mathcal{I}^{-1}$ is taken to be 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.

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

3. Bayesian Methods

Bayesian method is a different approach to estimating parameters of a model.

From the Bayes' Theorem: let A and B be two events, and $P(B) \neq 0$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5}$$

$$= \frac{P(B|A)P(A)}{P(B)} \tag{6}$$

Let $X_1, \ldots, X_n$ be i.i.d from the pdf $f(x|\theta)$, where θ is an unknown parameter. Let $\pi(\theta)$ denote the researcher's prior belief about θ. Now $\pi(\theta)$ is a pdf.

Then, given the realization $(x_1, \ldots, x_n)$ from the joint likelihood $f(x_1, \ldots, x_n|\theta)$, we *update* our prior according to the Bayes' rule:

$$\pi(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{f(x_1, \ldots, x_n)} \tag{7}$$

Where $f(x)$ is the marginal distribution of x, i.e. $f(x) = \int f(x|\theta)\pi(\theta)d\theta$.

$\pi(\theta|x_1, \ldots, x_n)$ is called the *Posterior distribution* of θ. Our *Bayes estimator* is $\pi(\theta|x_1, \ldots, x_n)$, which is an entire probability distribution, not a single point estimate. To report a single point estimate from this distribution, we usually report the mean of this posterior distribution, called the posterior mean, $\mathbb{E}_{\pi(\theta|x_1,\ldots,x_n)}[\hat{\theta}] = \int \theta\pi(\theta|x_1, \ldots, x_n)d\theta$.

To quantify the uncertainty around the posterior mean, we can report the posterior variance, which is: $\text{Var}_{\pi(\theta|x_1,\ldots,x_n)}(\hat{\theta})$. Alternatively, we can also report the mode, the median, or other summary statistics of the posterior distribution.

The marginal distribution $f(x)$ does not depend on θ, it is just a constant. As such, we can express the posterior distribution as:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \tag{8}$$

The constant of proportionality can be found by computing the constant C such that $Cf(x|\theta)\pi(\theta)$ integrates to one (with respect to θ), ensuring that $\pi(\theta|x)$ is a valid probability distribution.

3.1. Frequentist vs Bayesian estimators

All the estimators we encountered prior to this lecture have been Frequentist estimators. A *Frequentist* estimator of θ is a function of only the data $(x_1, \ldots, x_n)$. For instance, the sample mean and the sample variance are Frequentist estimators. The Maximum Likelihood estimator (MLE), $\hat{\theta}(x_1, \ldots, x_n) = \text{argmax}_\theta f(x_1, \ldots, x_n|\theta)$, is also a Frequentist estimator.

Frequentist estimators	Bayesian estimators
θ is a constant (there is a ground truth).	θ is not a constant, there is no ground truth. Fundamentally, θ is a random variable.
Requires the sampling model $f(x_1, \ldots, x_n\|\theta)$	Requires $f(x_1, \ldots, x_n\|\theta)$ and a prior distribution $\pi(\theta)$.
Given a random sample $(X_1, \ldots, X_n)$ from $f(x_1, \ldots, x_n\|\theta)$, estimate θ as a function of $(X_1, \ldots, X_n)$. Typically involved optimization.	Compute $\pi(\theta\|x_1, \ldots, x_n) = \frac{f(x_1,\ldots,x_n\|\theta)\pi(\theta)}{f(x)}$.
Uncertainty about the estimate is given by the sampling distribution.	Uncertainty about the estimate is given by the posterior distribution.
Variation in the sampling distribution is entirely due to the sampling variation in $f(x_1, \ldots, x_n\|\theta)$.	Variation in the posterior distribution is a combination of sampling distribution and the prior distribution.

3.2. Prior distribution

What is the prior distribution $\pi(\theta)$ and how do we specify it? It is a lengthy topic, but roughly:

(i) A prior distribution can be entirely subjective. A researcher's subjective belief about θ.

(ii) A prior distribution can be derived from other models and previous studies. As such, the posterior distribution reflects an updating of the prior $\pi(\theta)$ to the posterior $f(\theta|x)$ when confronted with the data x. In practice, Bayesian methods perform well because it is a form of model-averaging or data-combination.

(iii) A prior distribution itself can be estimated from the current data x. This is the Empirical Bayes approach.

(iv) A prior distribution reflects *model uncertainty*. For instance, you are not sure that $f(x|\theta,\sigma) \sim \mathcal{N}(\theta,\sigma)$ is the right model, so you let $\theta \sim \mathcal{N}(\mu,\tau)$, essentially considering a large class of Normal distributions with varying locations as the model. The prior distribution is $\mathcal{N}(\mu,\tau)$.[3]

4. Example

4.1. Normal distributions

Let $X_1,\ldots,X_n$ are iid $\sim \mathcal{N}(\theta,\sigma^2)$, and suppose that the prior distribution is $\pi(\theta) = \mathcal{N}(\mu,\tau^2)$. Just for this example, assume that τ,μ,σ are known.

The likelihood:

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\theta)^2/2\sigma^2}$$

The prior:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-(\theta-\mu)^2/2\tau^2}$$

The posterior:

$$f(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)\pi(\theta)}{\int f(x_1,\ldots,x_n|\theta)\pi(\theta)d\theta}$$

Observe that $\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_i^n (x_i-\bar{x})^2 + n(\bar{x}-\theta)^2$, where $\bar{x} = \frac{1}{n}\sum_i^n x_i$ is the sample mean.

$$\begin{aligned} f(\theta|x_1,\ldots,x_n) &\propto \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\right)\exp\left(\frac{-(\theta-\mu)^2}{2\tau^2}\right) \\ &\propto \exp\left(\frac{-(\theta-\tilde{\mu})^2}{2\tilde{\tau}^2}\right) \end{aligned}$$

Where:

[3] We can set μ to be the sample mean, in the spirit of Empirical Bayes. A reasonable prior would then be $\theta \sim \mathcal{N}(\frac{1}{n}\sum_{i=1}^n x_i, 1)$. Our Bayes estimate of σ would be robust to model misspecification.

$$\tilde{\mu} = \tilde{\tau}^2 \left(\frac{n}{\sigma^2} \bar{x} + \frac{1}{\tau^2} \mu \right)$$

$$\tilde{\tau}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}$$

Therefore, the posterior distribution of θ given $x_1, \ldots, x_n$ is:

$$\theta | x_1, \ldots, x_n \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\tau}^2\right) \tag{9}$$

Since the variance of the sample mean is σ^2/n, and the variance of the prior distribution is τ^2, then the variance of the posterior distribution is just a harmonic mean between the two variances, σ^2/n and τ^2.

The posterior mean is a weighted sum of the prior mean μ and the sample mean $\bar{x}$ with weights that reflect the precision of the sample mean (given by the reciprocal of σ^2/n) and the precision of the prior (given by the reciprocal of τ^2).

As the sample size n increases, the posterior mean becomes more similar to the sample mean, which is the Frequentist estimator. This means that information from the sample dominates the prior, and the variance of the posterior reflects mostly sampling uncertainty.

As a numerical example, suppose the prior is $\pi(\theta) = \mathcal{N}(5, 3)$. Suppose $\bar{x} = 10$, and the sample size is $n = 50$. The sampling distribution of the sample mean is $\mathcal{N}(10, \frac{\sigma^2}{50})$, suppose further that $\sigma^2 = 100$.

The posterior distribution is $\mathcal{N}(8, 1.2)$ according to (9). The frequentist approach would pick the sample mean $\bar{x} = 10$, as an estimator for θ.

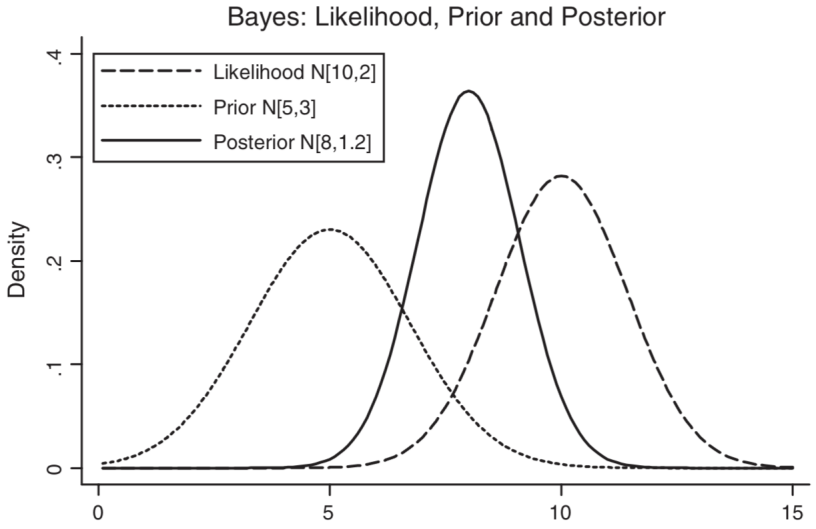

FIGURE 1. Figure from Cameron and Trivedi's "Microeconometrics: Methods and Applications"

5. Uninformative prior

Given the likelihood function $f(x|\theta)$, suppose the parameter range of θ is finite. Let the prior be a Uniform density over the range of θ, i.e. $\pi(\theta) = \frac{1}{C}$, for some number C.

The posterior distribution is:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$
$$\pi(\theta|x) \propto f(x|\theta)$$

Recall that MLE is $\text{argmax}_\theta f(x|\theta)$. Therefore, MLE coincides with the *mode of the posterior distribution* when the prior is uninformative!

However when the range of θ is unbounded, a Uniform prior distribution is not proper or well-defined. In practice, we often impose a diffuse prior or a flat prior $\mathcal{N}(0, 100)$, which is a very flat distribution. It is approximately $\pi(\theta) \approx 1/(2\pi \times 100)$, which is a constant.

6. Conjugate prior

The prior distribution $\pi(\theta)$ is a *Conjugate Prior* for the likelihood function $f(x|\theta)$ if the resulting posterior distribution $f(\theta|x)$ belongs to the same probability distribution family as the prior.

In the example before, the conjugate prior for a Normal distribution is a Normal distribution, since the posterior distribution is also a Normal distribution.

In the next example, the conjugate prior for Binomial(n, p) with p as the unknown is the Beta distribution.

6.1. Example: Binomial Bayes Estimator

Suppose $X|p \sim$ Binomial(n, p). Let the prior distribution be $\pi(p) \sim$ Beta(α, β). Then it turns out, the posterior distribution given the realization $X = x$ is $p|x \sim$ beta$(x + \alpha, n - x + \beta)$.

The mean of the prior distribution Beta(α, β), is $\frac{\alpha}{\alpha+\beta}$. The posterior mean is $\frac{x+\alpha}{n+\alpha+\beta}$, which can be decomposed into a weighted average between the sample information and the prior information:

$$\frac{x+\alpha}{n+\alpha+\beta} = \frac{n}{\alpha+\beta+n}\left(\frac{x}{n}\right) + \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)$$

Here, x/n is the frequentist estimator for p

6.2. MSE of Bayes Estimator

*Optional reading

Let $\hat{p}_B = \frac{x+\alpha}{n+\alpha+\beta}$ be the Bayes estimator of the Binomial parameter p.

The Mean Square Error of this Bayes estimator is:

$$\begin{aligned}
\mathbb{E}[(\hat{p}_B - p)^2] &= \text{Var}(\hat{p}_B) + (\mathbb{E}[\hat{p}_B] - p)^2 \\
&= \text{Var}\left(\frac{x+\alpha}{n+\alpha+\beta}\right) + \left(\mathbb{E}\left[\frac{x+\alpha}{n+\alpha+\beta}\right] - p\right)^2 \\
&= \frac{np(1-p)}{(n+\alpha+\beta)^2} + \left(\frac{np+\alpha}{n+\alpha+\beta} - p\right)^2
\end{aligned}$$

We could try to tune α and β to minimize the MSE. At the choice of $\alpha = \beta = \sqrt{n/4}$, the MSE is minimized and does not depend on p. (This is an attactive feature because regardless of what the true parameter is, the MSE is guarantee to be that number)

The MSE of the Bayes estimator $\hat{p}_B = \frac{x+\sqrt{n/4}}{n+\sqrt{n}}$ then becomes:

$$\mathbb{E}[(\hat{p}_B - p)^2] = \frac{n}{4(n+\sqrt{n})^2}$$

The frequentist (MLE) estimator of p is $\hat{p} = x/n$. The MSE is:

$$\mathbb{E}[(x/n - p)^2] = \text{Var}(x/n) + (\mathbb{E}[x/n] - p)^2 = \frac{p(1-p)}{n} + 0$$

Comparing the two MSEs, the Bayes estimator does better for an intermediate range of p and when n is small.

Is this surprising? The Bayes estimator is more flexible, and has more parameters that one can tune to optimize the MSE.

Essentially, we are considering a distribution over distributions, and in this sense, the estimator is more robust. The Bayes estimator $\hat{p}_B$ doesn't just estimate Binomial(n, p), but Binomial(n, p) over $p \sim \pi(\alpha, \beta)$.

7. More advanced Bayesian topics

7.1. Markov Chain Monte Carlo (MCMC)*

In practice, the posterior density has no closed form. If the posterior density is univariate, then we can use probability integral transform to sample from this posterior density. More generally, the posterior density will be multivariate. The most common way to sample from a multivariate density is to use the MCMC (Markov Chain Monte Carlo) method.

Essentially, the heavy-lifting in Bayesian analysis is sampling, while the heavy-lifting in frequentist analysis is optimization.[4]

[4]Bear in mind, optimization is usually parallelizable, while sampling is sequential. As such, bayesian estimation can be slower. However, there are new classes of bayesian methods are parallelizable.