Math 502AB - Lecture 22

Dr. Jamshidian

November 13, 2017

1 Lecture - Part 1

1.1 Prior and Posterior Distributions

Let X be a random variable with its distribution depending on $\vec{\theta}$. We assume that $\vec{\theta}$ is random and it has a distribution $f_\Theta(\vec{\theta})$. This distribution is called the **prior distribution**. It has this name mainly because it does not depend on data, but rather *prior experience*. We assume that:

$$X|\Theta \sim f_{X|\Theta}(x|\theta)$$

Let $X_1, ..., X_n \sim f_{X|\Theta}(x|\theta)$ (**iid**). Then:

1. The joint density of $\vec{X} = (X_1, ..., X_n)$ given $\Theta = \theta$ is given by:

$$L(X|\Theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta)$$

2. The joint density of $\vec{X}$ and Θ is:

$$f_{X,\Theta}(x, \theta) = L(X|\Theta = \theta) f_\Theta(\theta)$$

3. The marginal distribution of $\vec{X}$ is:

$$f_X(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(X|\Theta) f_\Theta(\theta) d\vec{\theta}$$

4. The posterior distribution of Θ given $\vec{X}$ is:

$$f_{\Theta|\vec{X}}(\theta|X) = \frac{L(\theta|X) f_\Theta(\theta)}{f_{\vec{X}}(x)}$$

   Posterior refers to *posterior to observing the data*. In other words it is an update about your knowledge of θ depending on the data observed. Any time you are making inference in the Bayesian world, you are really making estimations on the **posterior distribution**.

Example:

Consider the model $X_i|\Theta = \theta \sim Poisson(\theta)$. Consider:

$$\text{Prior: } \Theta \sim gamma(\alpha, \beta)$$

The posterior is determined by:

$$\begin{aligned}
f_{\Theta|X}(\theta|x) &\propto \mathcal{L}(\theta|X) f_\Theta(\theta) \\
&= \left(\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \right) \cdot \left(\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha} \right) \\
&= \left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) e^{-n\theta}\theta^{\sum x_i}\theta^{\alpha-1}e^{-\theta/\beta} \\
&\propto e^{-\theta\left(n+\frac{1}{\beta}\right)}\theta^{\sum x_i+\alpha-1}
\end{aligned}$$

Which is the *kernel of a gamma distribution*. So we have:

$$\Theta|X \sim gamma\left(\sum x_i + \alpha, \frac{\beta}{n\beta+1} \right)$$

We can see how the *prior* gets modified by the data. One thing to notice in this example, is that we started off with a prior that was *gamma*, and ended up with a posterior that was *gamma* as well. We describe the *gamma* as a **conjugate prior** of the *Poisson.*

1.2 Section 7.3 - Methods of Evaluating Estimators

1.2.1 Desired Properties of Estimators

1. **Unbiasedness:** An estimator W is an *unbiased estimate* of a parameter θ if $E[W] = \theta$. In general, the bias of an estimator is:

$$Bias_\theta(W) = E[W] - \theta$$

Example:

Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$.

$$\begin{aligned}
E[\overline{X}] &= \mu \quad \overline{X} \text{ is an unbiased estimator of } \mu \\
E[S^2] &= \sigma^2 \quad S^2 \text{ is an unbiased estimator of } \mu \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^62 \\
Bias_{\sigma^2}(\hat{\sigma}^2) &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\
&= -\frac{1}{n}\sigma^2
\end{aligned}$$

2. **Variance:** Suppose we have an estimator that is centered around θ. We might prefer the estimator with lower variance if all else is equal.

   Example:

$$\begin{aligned}
Var\left(S^2\right) &= Var\left[\left(\frac{(n-1)S^2}{\sigma}\right)\left(\frac{\sigma^2}{n-1}\right)\right] \\
&= \frac{\sigma^2}{(n-1)^2} \cdot Var\left(\chi^2_{(n-1)}\right) \\
&= \frac{2\sigma^4}{n-1} \\
Var(\hat{\sigma}^2) &= Var\left(\frac{n-1}{n}S^2\right) \\
&= \left(\frac{n-1}{n}\right)^2 Var(S^2) \\
&= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} \\
&= \frac{2(n-1)}{n^2}\sigma^4
\end{aligned}$$

3. **Mean Squared Error:** The *mean squared error* (**MSE**) of an estimator W for a parameter θ is defined by:

$$MSE_\theta(W) = E\left[W - \theta\right]^2$$

   Note:

$$\begin{aligned}
E[W-\theta]^2 &= E[W - E(W) + E(W) - \theta]^2 \\
&= E[W - E(W)]^2 + E[E(W) - \theta]^2 \\
&= Var(W) + \left(Bias_\theta(W)\right)^2
\end{aligned}$$

   Example:

$$\begin{aligned}
MSE(\overline{X}) &= Var(\overline{X}) = \frac{\sigma^2}{n} \\
MSE(S^2) &= Var(S^2) = \frac{2\sigma^4}{n-1} \\
MSE(\hat{\sigma}^2) &= Var(\hat{\sigma}^2) + \left(Bias_{\sigma^2}(\hat{\sigma}^2)\right)^2 \\
&= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}
\end{aligned}$$

Example: Suppose $X_1, ..., X_n \sim Bernoulli(p)$ the MLE for p is given by $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If we assume a prior $P \sim Beta(\alpha, \beta)$ then a Baye's point estimate is

$$\hat{p}_B = \frac{\sum x_i + \alpha}{\alpha + \beta + n}$$

We wish to compare the MSE for $\hat{p}$ and $\hat{p}_B$:

$$MSE(\hat{p}) = Var(\hat{p}) = \frac{p(1-p)}{n}$$

We have:

$$\begin{aligned}
Var\left(\hat{p}_B\right) &= \left(\frac{1}{\alpha+\beta+n}\right)^2 Var\left(\sum_{i=1}^{n} x_i\right) = \frac{np(1-p)}{\alpha+\beta+n} \\
Bias_P(\hat{p}_B) &= E\left[\frac{\sum x_i + \alpha}{\alpha+\beta+n}\right] - p \\
&= \frac{np+\alpha}{\alpha+\beta+n} - p \\
MSE(\hat{p}_B) &= \frac{np(1-p)}{\alpha+\beta+n} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2 \\
&= \frac{p^2((\alpha+\beta)^2 - n) + p[n - 2\alpha(\alpha+\beta) + \alpha^2]}{(\alpha+\beta+n)^2}
\end{aligned}$$

We wish to show that depending on different values of p and n, we might want to use one estimator over the other.

Consider a special case where $n = (\alpha+\beta)^2$, $n - 2\alpha(\alpha+\beta) = 0$, $\alpha = \sqrt{n}/2$, and $\beta = \sqrt{n}/2$. Then we have:

$$\begin{aligned}
MSE(\hat{p}_B) &= \frac{n}{4(n+\sqrt{n})^2} \\
MSE(\hat{p}) &= \frac{p(1-p)}{n}
\end{aligned}$$

2 Lecture - Part 2

2.1 Section 7.3.2 - Best Unbiased Estimator

Definition:

An estimator W^* is a **best unbiased estimator** of θ if for all θ and any other unbiased estimator W, it satisfies

$$Var(W^*) \leq Var(W)$$

W^* is also called a **uniform minimum variance** unbiased estimator (UMVUE) of θ.

2.1.1 Theorem: Cramer-Rao Inequality (Lower Bound)

Let $X_1, ..., X_n$ be a sample from *pdf* $f(x|\theta)$ and $W(\vec{X}) = W(X_1, ..., X_n)$ be an estimator satisfying:

$$\frac{d}{d\theta} E\left[W(\vec{X})\right] = \int_X \cdots \int_X \frac{\partial}{\partial\theta} W(\vec{X}) f(X|\theta) dx_1 \cdots dx_n$$

and $Var(W(X)) < \infty$. Then:

$$Var(W(X)) \geq \frac{\left(\frac{d}{d\theta} E[W(X)]\right)^2}{E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2} \quad (**)$$

Proof:

The proof is an application of the Cauchy-Schwartz Inequality.

$$(Cov(X,Y))^2 \leq Var(X)Var(Y)$$
$$\Rightarrow Var(X) \geq \frac{[Cov(X,Y)]^2}{Var(Y)}$$

Consider:

$$\begin{aligned}
E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right] &= \int \cdots \int \left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right) f(X|\theta) dx_1 \cdots dx_n \\
&= \int \cdots \int \frac{f'(X|\theta)}{f(X|\theta)} f(X|\theta) dx_1 \cdots dx_n \\
&= \frac{d}{d\theta} \int \cdots \int f(X|\theta) dx_1 \cdots dx_n = \frac{d}{d\theta} 1 = 0
\end{aligned}$$

This suggests letting $Y = \frac{\partial}{\partial\theta} \log f(X|\theta)$ and letting X to be $W(X)$ in $(**)$ [the *Cauchy-Schwarz Inequality*]. It is sufficient to show that:

$$Cov\left(W(X), \frac{\partial}{\partial\theta} \log f(X|\theta)\right) = \frac{\partial}{\partial\theta} E[W(X)]$$

So we take:

$$\begin{aligned}
Cov\left(W(X), \frac{\partial}{\partial\theta} \log f(X|\theta)\right) &= E\left[W(X) \cdot \frac{\partial}{\partial\theta} \log f(X|\theta)\right] - E[W(X)] E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right] \\
&= E\left[W(X) \frac{\frac{\partial}{\partial\theta} f(X|\theta)}{f(X|\theta)}\right] \\
&= \int \cdots \int \frac{W(X) \frac{\partial}{\partial\theta} f(X|\theta)}{f(X|\theta)} f(X|\theta) dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial\theta} \int \cdots \int W(X) f(X|\theta) dx_1 \cdots dx_n = \frac{\partial}{\partial\theta} E[W(X)]
\end{aligned}$$

2.1.2 CRLB for iid Case

If $X_1, \ldots, X_n \sim f(X|\theta)$ (**iid**), then:

$$Var(W(X)) \geq \frac{\left[\frac{d}{d\theta} E[W(X)]\right]^2}{nE\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2}$$

So where does this n term come from? Well, we have:

$$\begin{aligned} E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2 &= E\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)\right]^2 \\ &= \sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right]^2 + \sum_{i\neq j}\sum E\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\frac{\partial}{\partial\theta} \log f(X_j|\theta)\right] \end{aligned}$$

But we know that $E\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right] = 0$, so the double sum is eliminated. Thus we have the equality of $nE\left[\frac{\partial}{\partial n} \log f(X|\theta)\right]$, as wanted.

2.1.3 Fisher Information

The quantity:

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

is known as the **Fisher information**. Under some *regularity conditions*:

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

Example

Let $X_1, ..., X_n \sim Poisson(\lambda)$. We know that $E[\overline{X}] = \lambda$, and $E[S^2] = \lambda$. So we have two estimators which are unbiased. Our question then becomes, *which one is better.*

Since they are unbiased, the one that has the smaller variance will be our answer. We know:

$$Var(\overline{X}) = \frac{\lambda}{n}$$

But the variance for S^2 is complicated! So is there a way to just say $\overline{X}$ is better? Well, if we show that the variance above is equal to the *Cramer-Rao Lower Bound*, then we have it!

The *Cramer-Rao Lower Bound* is:

$$CRLB = \frac{\left(\frac{d}{d\theta} E[W(X)]\right)^2}{n\left(E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]\right)^2}$$

We know the numerator is equal to 1. So let's compute the denominator:

$$\begin{aligned}
\log f(X|\lambda) &= \log\left(\frac{e^{-\lambda}\lambda^x}{x!}\right) = -\lambda + x\log\lambda - \log x! \\
\frac{\partial}{\partial\lambda}\log f(X|\lambda) &= -1 + \frac{x}{\lambda} \\
\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda) &= -\frac{x}{\lambda^2} \\
\mathcal{I}(\lambda) &= -E\left[\frac{-X}{\lambda^2}\right] = \frac{1}{\lambda} \\
\Rightarrow CRLB &= \frac{1}{n\cdot\frac{1}{\lambda}} = \frac{\lambda}{n}
\end{aligned}$$

Therefore $\overline{X}$ is the best unbiased estimator for λ.

Example:

Let $X_1, \ldots, X_n \sim f(X|\theta)$ (**iid**), where:

$$f(X|\theta) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

We then have:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f(X|\theta) &= -\frac{1}{\theta} \\
E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 &= \frac{1}{\sigma^2}
\end{aligned}$$

If we were to use the Cramer-Rao Lower Bound, then:

$$Var(W(X)) \geq \frac{\theta^2}{n}$$

Consider the estimator $W(X) = X_{(n)}$. We have:

$$\begin{aligned}
E[W] &= \int_0^\theta \frac{w^n n}{\theta^n} dw = \frac{n}{n+1}\theta \\
E\left[\frac{n+1}{n}W\right] &= \theta \quad \text{, an unbiased estimator}
\end{aligned}$$

If you calculate the *variance* of this estimator, you get:

$$\begin{aligned}
Var\left(\frac{n+1}{n}W\right) &= \left(\frac{n+1}{n}\right)^2 Var(W) \\
&= \frac{1}{n(n+2)}\theta^2 < \frac{\theta^2}{n}
\end{aligned}$$