Bayes Estimation

February 26-28, 2008

Bayesian estimation begins with an assumed probability distribution on the parameter space Θ. With this approach θ itself is a random variable and the observations $X_1, \ldots, X_n$ are conditionally independent given the value of θ. Consequently, in Bayesian statistics, (X, θ) is a random variable on the state space $S^n \times \Theta$.

The distribution of θ on Θ is called the **prior distribution**. We shall denote its density by π. Together, the prior distribution and the parametric family $\{P_\theta; \theta \in \Theta\}$ determine the joint distribution of (X, θ).

1 Bayes Formula

For events A and B, recall that the conditional probability is

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A),$$

or

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

Now, if we set

$$A = \{\theta = \theta_0\} \quad \text{and} \quad B = \{X = \mathbf{x}\},$$

then

$$P\{\theta = \theta_0 | X = \mathbf{x}\} = \frac{P\{X = \mathbf{x}|\theta = \theta_0\}P\{\theta = \theta_0\}}{P\{X = \mathbf{x}\}}.$$

If the appropriate densities exist, then we can write **Bayes formula** as

$$f_{\Theta|X}(\theta_0|\mathbf{x}) = \left(\frac{f_{X|\Theta}(\mathbf{x}|\theta_0)}{\int f_{X|\Theta}(\mathbf{x}|\tilde{\theta})\pi(\tilde{\theta})\, d\tilde{\theta}} \right) \pi(\theta_0),$$

to compute the **posterior density** $f_{\Theta|X}(\theta_0|x)$ as the product of the **Bayes factor** and the **prior density**.

If T is a sufficient statistic and $f_{X|\Theta}(\mathbf{x}|\tilde{\theta}) = h(\mathbf{x})g(\theta, T(\mathbf{x}))$, then the Bayes factor

$$\frac{f_{X|\Theta}(\mathbf{x}|\theta_0)}{\int f_{X|\Theta}(\mathbf{x}|\tilde{\theta})f_\Theta(\tilde{\theta})\, d\tilde{\theta}} = \frac{h(\mathbf{x})g(\theta, T(\mathbf{x}))}{\int h(\mathbf{x})g(\tilde{\theta}, T(\mathbf{x}))\, d\tilde{\theta}} = \frac{g(\theta, T(\mathbf{x}))}{\int g(\tilde{\theta}, T(\mathbf{x}))\, d\tilde{\theta}}$$

is a function of T.

Example 1 (Normal observations and normal prior)**.** *Suppose that*

- *θ is $N(\theta_0, 1/\lambda)$, and*
- *that given θ, X consists of n conditionally independent $N(\theta, 1)$ random variables.*

Then the prior density is $f_\Theta(\theta) = \sqrt{\frac{\lambda}{2\pi}} \exp(-\frac{\lambda}{2}(\theta - \theta_0)^2)$, *and*

$$\begin{aligned} f_{X|\Theta}(\mathbf{x}|\theta) &= (2\pi)^{-n/2} \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2) \\ &= (2\pi)^{-n/2} \exp(-\frac{n}{2}(\theta - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2). \end{aligned}$$

The posterior density is proportional to

$$k(\mathbf{x}) \exp(-\frac{1}{2}(n(\theta - \bar{x})^2 + \lambda(\theta - \theta_0)^2))$$

$$= \tilde{k}(\mathbf{x}) \exp(-\frac{n+\lambda}{2}(\theta - \tilde{\theta}(\mathbf{x}))^2).$$

where

$$\tilde{\theta}(\mathbf{x}) = \frac{\lambda\theta_0 + n\bar{x}}{\lambda + n}.$$

Thus, the posterior distribution is

$$N(\theta_1(\mathbf{x}), 1/(\lambda + n)).$$

Note that it is a function of the sufficient statistics $T(\mathbf{x}) = x_1 + \cdots + x_n$. If n is small, then $\theta_1(\mathbf{x})$ is near θ_0. If n is large, $\theta_1(\mathbf{x})$ is near $\bar{x}$.

2 Bayes Action

Recall that given a loss function $\mathcal{L}$ and an estimator d the risk function $\mathcal{R} : \Theta \times \mathcal{D} \to \mathbb{R}$ is the expected loss for that decision.

$$\mathcal{R}(\theta, d) = E_\theta \mathcal{L}(\theta, d(X))$$

and the mean risk, or **Bayes risk**,

$$r(\pi, d) = \int_\Theta \mathcal{R}(\theta, d)\pi(\theta)\ d\theta = \int_\Theta \int_{\mathbb{R}^n} \mathcal{L}(\theta, d(x)) f_X(\mathbf{x}|\theta)\pi(\theta)\ d\mathbf{x}d\theta.$$

The decision function that minimizes risk is called the **Bayes action**.

If the loss function is $\mathcal{L}_1(\theta, a) = |\theta - a|$, then the posterior median minimizes risk and thus the Bayes action $\hat{\theta}_1(\mathbf{x})$ satisfies

$$\frac{1}{2} = \int_{-\infty}^{\hat{\theta}_1(\mathbf{x})} f_{\Theta|X}(\theta|x)\ d\theta.$$

If the loss function is $\mathcal{L}_2(\theta, a) = (\theta - a)^2$, then the posterior mean minimizes risk and thus the Bayes action

$$\hat{\theta}_2(\mathbf{x}) = E[\theta|X = \mathbf{x}] = \int \theta f_{\Theta|X}(\theta|\mathbf{x})\ d\theta.$$

For the example of a normal prior and normal observations, $\hat{\theta}_1(\mathbf{x}) = \hat{\theta}_2(\mathbf{x}) = \tilde{\theta}(\mathbf{x})$.

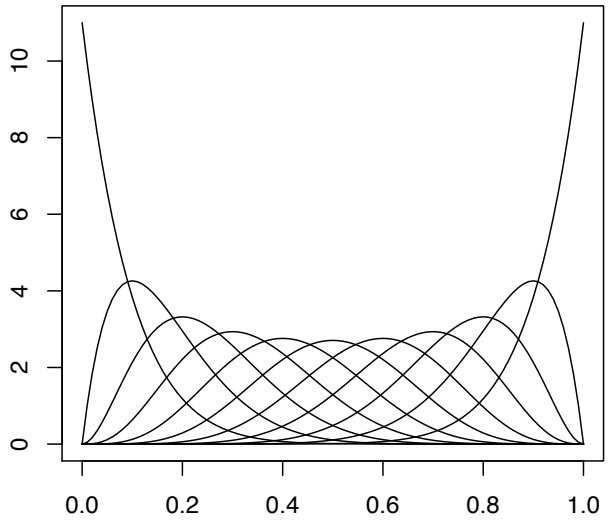

Figure 1: Beta posterior distribution with $t = 0, 1, \cdots, 10$ successes in 10 Bernoulli trials based on a uniform prior

Example 2. *Let the prior distribution π on θ be a beta distribution with parameters α and β and consider Bernoulli observations $X_1, \ldots, X_n$ with parameter θ. $T(X) = X_1 + \cdots + X_n$ is a sufficient statistic. The posterior distribution*

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \mathbf{L}(\theta|\mathbf{x})\pi(\theta) = \theta^{T(\mathbf{x})}(1-\theta)^{n-T(\mathbf{x})}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 \le \theta \le 1.$$

Thus,

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto \theta^{T(\mathbf{x})+\alpha-1}(1-\theta)^{n-T(\mathbf{x})+\beta-1} \quad 0 \le \theta \le 1.$$

and the posterior distribution is $Beta(T(\mathbf{x}) + \alpha, n - T(\mathbf{x}) + \beta)$. If we want to estimate θ using a quadratic risk function, then

$$\hat{\theta}(\mathbf{x}) = E[\theta|X = \mathbf{x}] = \frac{T(\mathbf{x}) + \alpha}{n + \alpha + \beta}.$$

The uniform distribution on $[0, 1]$ has a $Beta(1, 1)$ distribution. In this case

$$\hat{\theta}(\mathbf{x}) = \frac{T(\mathbf{x}) + 1}{n + 2}.$$

The posterior densities are graph using R *in Figure 1 using*

```
> curve(dbeta(x,1,11),0,1)
> for (i in 2:11){curve(dbeta(x,i,12-i),0,1,add=TRUE)}
```