Lecture 06: Bayesian Estimation

Mathematical Statistics II, MATH 60062/70062

Tuesday February 8, 2022

Reference: Casella & Berger, 7.2.3

Statistical inference

- Frequentist/classical approach
  - Treat model parameter θ as *fixed* (and unknown).
  - Observe $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$.
  - Use $\boldsymbol{x}$ to make inference about θ.
- Bayesian approach
  - Model $\theta \sim \pi(\theta)$. So θ is considered as a *random* quantity whose variation is described by $\pi(\theta)$.
  - Observe $\boldsymbol{X} \mid \theta \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$.
  - Update $\pi(\theta)$ with $f(\theta \mid \boldsymbol{x})$.

Bayesian inference

Bayesian inference refers to the updating of prior beliefs into posterior beliefs conditional on observed data using **Bayes' Theorem**,

$$f(\theta \mid \boldsymbol{x}) = \frac{f_{\boldsymbol{X},\theta}(\boldsymbol{x}, \theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int_{\Theta} f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta}.$$

- The **prior distribution** $\pi(\theta)$ describes our belief that θ represents the true population characteristics (not related to information provided by the data $\boldsymbol{x}$).
- The **sampling model (likelihood)** $f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)$ describes our belief that $\boldsymbol{x}$ would be the outcome if we knew θ to be true.
- The **posterior distribution** $f(\theta \mid \boldsymbol{x})$ describes our belief that θ is the true value, having observed the sample $\boldsymbol{x}$.

Bernoulli Bayesian inference

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(\theta)$, where the prior distribution on θ is $\mathrm{Beta}(a, b)$, and the values of a and b are known.

- The **prior distribution** of θ,

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

- The **sampling model (likelihood)**,

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

- The **joint distribution** of $\boldsymbol{X}$ and θ,

$$\begin{aligned} f_{\boldsymbol{X},\theta}(\boldsymbol{x}, \theta) &= f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{\sum_{i=1}^{n} x_i + a - 1}(1-\theta)^{n-\sum_{i=1}^{n} x_i + b - 1}. \end{aligned}$$

- The **marginal distribution** of $\boldsymbol{X}$,

$$
\begin{aligned}
f_{\boldsymbol{X}}(\boldsymbol{x}) &= \int_0^1 f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(\sum_{i=1}^n x_i + a)\Gamma(n - \sum_{i=1}^n x_i + b)}{\Gamma(n+a+b)}.
\end{aligned}
$$

- The **posterior distribution** of θ given $\boldsymbol{X} = \boldsymbol{x}$,

$$
\begin{aligned}
f(\theta \mid \boldsymbol{x}) &= \frac{f_{\boldsymbol{X},\theta}(\boldsymbol{x}, \theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} \\
&= \frac{\Gamma(n+a+b)}{\Gamma(\sum_{i=1}^n x_i + a)\Gamma(n - \sum_{i=1}^n x_i + b)} \theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^n x_i + b - 1}.
\end{aligned}
$$

The **posterior distribution** of θ is $\mathrm{Beta}(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b)$.

Here we can skip calculating the marginal distribution, which does not depend on θ.

$$\begin{aligned} f(\theta \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta) \\ &= \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{free of } \theta} \underbrace{\theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^n x_i + b - 1}}_{\text{Beta kernel}} \end{aligned}$$

By recognizing the kernel of Beta distribution, we can conclude that the posterior is $\mathrm{Beta}(\sum_{i=1}^n x_i + a, n - \sum_{i=1}^n x_i + b)$.

The posterior distribution of θ depends on

- The parameters of the prior distributions a, b (i.e., the **hyperparameters**). In this example, a and b are also known as "pseudo-counts".
- The data $\boldsymbol{x}$ through the sufficient statistic $t(\boldsymbol{x}) = \sum_{i=1}^n x_i$.

Sufficient statistics in Bayesian inference

It is not a coincidence that the posterior depends on the data through the sufficient statistics.

If $T = T(\boldsymbol{X})$ is sufficient, we know (by the Factorization Theorem)

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta) = g(t \mid \theta)h(\boldsymbol{x}).$$

Therefore, the posterior distribution

$$\begin{aligned} f(\theta \mid \boldsymbol{x}) &\propto f_{\boldsymbol{X}}(\boldsymbol{x} \mid \theta)\pi(\theta) \\ &\propto g(t \mid \theta)\pi(\theta). \end{aligned}$$

This shows that the posterior will depend on the data $\boldsymbol{x}$ through the value of the sufficient statistic $t = T(\boldsymbol{x})$. We can therefore write the posterior distribution as depending on t only; i.e.,

$$f(\theta \mid t) \propto f_T(t \mid \theta)\pi(\theta).$$

Binomial Bayesian inference

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{Bern}(\theta)$, where the prior distribution on θ is $\mathrm{Beta}(a, b)$, and the values of a and b are known.

We know $T = T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is sufficient and $T \sim \mathrm{Bin}(n, \theta)$. The posterior distribution

$$\begin{aligned} f(\theta \mid \boldsymbol{x}) &\propto f_T(t \mid \theta)\pi(\theta) \\ &= \binom{n}{t}\theta^t(1-\theta)^{n-t}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &= \binom{n}{t}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{t+a-1}(1-\theta)^{n-t+b-1}. \end{aligned}$$

By recognizing the Beta kernel, we can conclude that the posterior is $\mathrm{Beta}(t+a, n-t+b)$.

Binomial Bayes estimator

With the posterior distribution, $\mathrm{Beta}(t+a, n-t+b)$, a natural estimate for θ is the **posterior mean**

$$\begin{aligned}\hat{\theta}_{\mathrm{B}} &= \frac{t+a}{n+a+b} \\ &= \left(\frac{n}{n+a+b}\right) \underbrace{\left(\frac{t}{n}\right)}_{\text{sample mean}} + \left(\frac{a+b}{n+a+b}\right) \underbrace{\left(\frac{a}{a+b}\right)}_{\text{prior mean}}\end{aligned}$$

Thus $\hat{\theta}_{\mathrm{B}}$ is a linear combination of the prior mean and the sample mean, with the weights being determined by a, b ("pseudo-counts"), and n (sample size).

Point estimators with posterior distribution

Posterior mean is not the only possible point estimator with the posterior distribution. Other possibilities include **posterior median** and **posterior mode**.

Comparing and choosing among these estimators requires us to discuss **loss functions** (CB, Sec. 7.3.4).

Conjugate family

Let $\mathcal{F} = \{f_X(x \mid \theta) : \theta \in \Theta\}$ denote the class of PDFs or PMFs. A class Π of prior distributions is a **conjugate family** for $\mathcal{F}$ if the posterior distribution is also in class Π.

- The Beta family is conjugate for the Binomial family.
- The Gamma family is conjugate for the Poisson family.
- ...

Normal Bayesian inference

Suppose that $X \sim \mathcal{N}(\theta, \sigma^2)$, where the prior distribution on θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ, and τ^2 are all known, then the posterior distribution of θ is a Normal, with mean and variance given by

$$\begin{aligned} E(\theta \mid x) &= \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu, \\ \mathrm{Var}(\theta \mid x) &= \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \end{aligned}$$

Normal Bayesian inference

Suppose that $X \sim \mathcal{N}(\theta, \sigma^2)$, where the prior distribution on θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ, and τ^2 are all known, then the posterior distribution of θ is a Normal, with mean and variance given by

$$
\begin{aligned}
E(\theta \mid x) &= \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu, \\
\mathrm{Var}(\theta \mid x) &= \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}.
\end{aligned}
$$

The Bayes estimator is, again, a linear combination of the prior and sample means.

- If the prior information is good, so that $\sigma^2 > \tau^2$, then more weight is given to the prior mean.
- As the prior information becomes more vague (i.e., $\tau^2 \uparrow$), the Bayes estimator gives more weight to the sample information.

Normal Bayesian inference

Suppose that $X \sim \mathcal{N}(\theta, \sigma^2)$, where the prior distribution on θ is $\mathcal{N}(\mu, \tau^2)$. Assuming that σ^2, μ, and τ^2 are all known, then the posterior distribution of θ is a Normal, with mean and variance given by

$$\begin{aligned} E(\theta \mid x) &= \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu, \\ \mathrm{Var}(\theta \mid x) &= \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}. \end{aligned}$$

The Bayes estimator is, again, a linear combination of the prior and sample means.

- If the prior information is good, so that $\sigma^2 > \tau^2$, then more weight is given to the prior mean.
- As the prior information becomes more vague (i.e., $\tau^2 \uparrow$), the Bayes estimator gives more weight to the sample information.
- What if, a larger sample from $\mathcal{N}(\theta, \sigma^2)$ is made available?