Chapter 7
Point Estimation
Bayes Estimation

Outline

Overview

Given data $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}$, we must make a decision, a choice from the action space, $\mathcal{A}$. Thus, we introduce the decision function or rule.

$$d : \mathcal{X} \to \mathcal{A}.$$

In parametric estimation, the action space is simply the parameter space Θ

Decisions have consequences, a measure of how seriously we view incorrect decisions. This leads to the introduction of the loss function,

$$\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}.$$

Thus, if the state of nature is θ, then $\mathcal{L}(\theta, a)$ is the loss incurred upon taking the action a.

Loss Functions and Risk

1. $\mathcal{L}_1(\theta, a) = |a - \theta|$,
2. $\mathcal{L}_2(\theta, a) = (a - \theta)^2$,
3. $\mathcal{L}_\infty(\theta, a) = 0$ if $\theta = a$ and $\mathcal{L}(\theta, a) = 1$ if $\theta \neq a$.

The goal is to make the choice from the set of decision functions that minimizes the mean loss.

Definition. Let $\mathcal{D}$ denote the collection of decision rules. The risk function

$$\mathcal{R} : \Theta \times \mathcal{D} \to \mathbb{R}$$

is defined by

$$\mathcal{R}(\theta, d) = E_\theta \mathcal{L}(\theta, d(X_1, \ldots, X_n)).$$

Loss Functions and Risk

Consider a single discrete random variable with mass function $p(\cdot|\theta)$ and decision function $d(x) = x$. Then,

$$\begin{aligned}\mathcal{R}_1(\theta, d) &= E_\theta \mathcal{L}_1(\theta, d(X)) = E|X - \theta| = \sum_x |x - \theta| p_X(x|\theta) \\ &= \sum_{x<\theta} (\theta - x) p_X(x|\theta) + \sum_{x \geq \theta} (x - \theta) p_X(x|\theta) \\ &= \theta P_\theta\{X < \theta\} - \theta P_\theta\{X \geq \theta\} - \sum_{x<\theta} x p_X(x|\theta) + \sum_{x \geq \theta} x p_X(x|\theta).\end{aligned}$$

$\mathcal{R}_1$ is a continuous piecewise linear function of θ with slope

$$P\{X < \theta\} - P\{X \geq \theta\} = 1 - 2P\{X \geq \theta\}.$$

Thus, $\mathcal{R}_1$ is decreasing if $P\{X \geq \theta\} > 1/2$ and increasing if $P\{X \geq \theta\} < 1/2$.

Consequently, $\mathcal{R}_1$ is minimized by taking θ equal to the median.

Loss Functions and Risk

$$\mathcal{R}_2(\theta, d) = E_\theta \mathcal{L}_2(\theta, d(X)) = E(X-\theta)^2 = \sum_x (x-\theta)^2 p_X(x|\theta)$$

Thus,

$$\frac{\partial}{\partial \theta}\mathcal{R}_2(\theta, d) = -\sum_x (x-\theta) p_X(x|\theta) = -EX + \theta.$$

Thus, the minimum is achieved by taking θ equal to the mean.

$$\mathcal{R}_\infty(\theta, d) = E_\theta \mathcal{L}_\infty(\theta, d(X)) = 0 \cdot P\{X = \theta\} + 1 \cdot P\{X \neq \theta\} = 1 - P\{X = \theta\}.$$

This is minimized by taking θ equal to the mode.

Bayesian Formula for Densities

In Bayesian statistics, (X, Ψ) is a random variable on the cross product of the state space and the parameter space. The density π of Ψ on the parameter space Θ is called the prior density. Thus, the prior density and the family $\{P_\theta; \theta \in \Theta\}$ give a hierarchical model that determines the joint distribution of (X, Ψ).

In this approach, both the parameter and the data are modeled as random. Inference is based on posterior density derived from Bayes formula

$$f_{\Theta|X}(\theta|\mathbf{x}) = \frac{f_{X,\Theta}(\mathbf{x}, \theta)}{f_X(\mathbf{x})} = \frac{f_X(\mathbf{x}|\theta)\pi(\theta)}{f_X(\mathbf{x})}.$$

where the denominator is the continuous mixture

$$f_X(\mathbf{x}) = \int_\Theta f_{X,\Theta}(\mathbf{x}, \theta)\ d\theta = \int_\Theta f_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)\ d\theta$$

Bayesian Formula for Densities

We can write Bayes formula

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

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

If $T(\mathbf{X})$ is a sufficient statistic, the factorization theorem guarantees that the density has the form $f_{X|\Theta}(\mathbf{x}|\tilde{\theta}) = h(\mathbf{x})g(\theta, T(\mathbf{x}))$, 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 $T(\mathbf{X})$.

Bayes Risk

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

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

The Bayes risk is the mean risk with respect to the prior density.

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

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

Bayes Action

If the loss function is $\mathcal{L}_1(\theta, a) = |\theta - a|$, then the losterior 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

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

Sequencing Updating

The Bayesian approach is amenable to sequential updating. For example, if we collect independent data in three batches, say $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, then the density for the entire data set $\mathbf{x}$ can be written

$$f_{X|\Theta}(\mathbf{x}|\theta) = f_{X_3|\Theta}(\mathbf{x}_3|\theta) \cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta) \cdot f_{X_1|\Theta}(\mathbf{x}_1|\theta).$$

To set the notation, write

- $X = (X_1, X_2, X_3)$ for the the sequential sets of random variables associated to the observations,
- $f_{\Theta|X_1}(\theta|\mathbf{x}_1)$ for the posterior density based on the data $\mathbf{x}_1$, and
- $f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1, \mathbf{x}_2)$ for the posterior density based on the data $(\mathbf{x}_1, \mathbf{x}_2)$,

Sequencing Updating

Then, the posterior density

$$\begin{aligned} f_{\Theta|X}(\mathbf{x}\theta) &\propto f_{X|\Theta}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3|\theta)\pi(\theta) = f_{X_3|\Theta}(\mathbf{x}_3|\theta) \cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta) \cdot f_{X_1|\Theta}(\mathbf{x}_1|\theta)\pi(\theta) \\ &\propto f_{X_3|\Theta}(\mathbf{x}_3|\theta) \cdot f_{X_2|\Theta}(\mathbf{x}_2|\theta) \cdot f_{\Theta|X_1}(\theta|\mathbf{x}_1) \\ &\propto f_{X_3|\Theta}(\mathbf{x}_3|\theta) \cdot f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1, \mathbf{x}_2) \end{aligned}$$

Thus,

- The posterior density $f_{\Theta|X_1}(\theta|\mathbf{x}_1) \propto f_{X_1|\Theta}(\mathbf{x}_1|\theta)\pi(\theta)$ serves are the prior density for $(\mathbf{x}_2, \mathbf{x}_3)$.
- The posterior density $f_{\Theta|X_1,X_2}(\theta|\mathbf{x}_1, \mathbf{x}_2) \propto f_{X_2|\Theta}(\mathbf{x}_2|\theta) \cdot f_{\Theta|X_1}(\theta|\mathbf{x}_1)$ serves are the prior density for $\mathbf{x}_3$.

Of course, this strategy can be used for any number of sequential updates.

Bayesian Statistics

Recall the example of a normal prior on Ψ of normal observations X. We take

- The prior density to be $N(\theta_1, 1/\lambda_0)$
- The observations $X_1, . , . , X_n$ are independent $N(\theta, \sigma^2)$
- Their mean $\bar{X} \sim N(\theta, \sigma^2/n)$
- The posterior distribution is $N(\theta_1(\bar{x}), \sigma^2/(n + \lambda_0\sigma^2))$ where

$$\theta_1(\bar{x}) = \frac{\lambda_0}{\lambda_0 + n/\sigma^2}\theta_1 + \frac{n/\sigma^2}{\lambda_0 + n/\sigma^2}\bar{x}.$$

Sequencing Updating

The posterior distribution has mean

$$\theta_1(\bar{x}) = \frac{\lambda_0}{\lambda_0 + n/\sigma^2}\theta_1 + \frac{n/\sigma^2}{\lambda_0 + n/\sigma^2}\bar{x}.$$

For the information, we have the transformation $\lambda \mapsto n + \lambda$ for n observations. With these two ideas, we compute the sequential updates.

prior		statistics		posterior	
$\lambda = 1/4,$ $\sigma^2 = 4$	$\mu = 0,$	$n = 6,$	$\bar{x} = 1.216$	$\lambda = 25/4$	$\mu = 24/25 \cdot 1.216$ $\mu = 1.16736$
$\lambda = 25/4$ $\sigma^2 = 4/25,$	$\mu = 1.16736$	$n = 3,$	$\bar{x} = 1.911$	$\lambda = 37/4$	$\mu = 25/37 \cdot 1.16736 + 12/37 \cdot 1.911$ $\mu = 1.408541$
$\lambda = 37/4$ $\sigma^2 = 4/37$	$\mu = 1.408541$	$n = 3,$	$\bar{x} = 0.811$	$\lambda = 49/4$	$\mu = 37/49 \cdot 1.408541 + 12/49 \cdot 0.811$ $\mu = 1.262204$

Sequencing Updating

To accomplish this in one step, note that

$$n = 6 + 3 + 3 = 12, \qquad \bar{x} = \frac{1}{12}(6 \cdot 1.216 + 3 \cdot 1.911 + 3 \cdot 0.811) = 1.2885.$$

prior	statistics	posterior
$\lambda = 1/4 \quad \mu = 0$ $\sigma^2 = 4$	$n = 12, \quad \bar{x} = 1.2885$	$\lambda = 49/4, \quad \mu = 48/49 \cdot 1.2885 = 1.262204$

Conjugate Family

Definition. Let $\mathcal{F}$ denote a class of density functions. A class $\mathcal{P}$ of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class $\mathcal{P}$ for all $\mathbf{f} \in \mathcal{F}$, all $\pi \in \mathcal{P}$, and all data $\mathbf{x} \in \mathcal{X}$ derived from observations from a given $\mathbf{f} \in \mathcal{F}$.

If the prior and posteror families can be indexed by parameter sets $\Theta_{\mathcal{P}}$ and $\Theta_{\mathcal{F}}$, respectively, then the data $\mathbf{x} \in \mathcal{X}$ induces a mapping

$$\Theta_{\mathcal{P}} \xrightarrow{\mathbf{x}} \Theta_{\mathcal{P}}$$

So, if $\mathcal{F} = Bin(n, p)$, n known, then $\mathcal{P} = Beta(\alpha, \beta)$ is a conjugate family for $\mathcal{F}$. If $T(\mathbf{x})$ is thenumber of success in n trials, then

$$(\alpha, \beta) \xrightarrow{\mathbf{x}} (\alpha + T(\mathbf{x}), \beta + n - T(\mathbf{x})).$$

If $\mathcal{F} = N(\theta, \sigma^2)$, σ^2 known, then $\mathcal{P} = N(\theta_1, \sigma_1^2)$ is a conjugate family for $\mathcal{F}$. If $\bar{x}$ is the mean for n observations, then

$$(\theta_1, \frac{1}{\lambda_0}) \xrightarrow{\mathbf{x}} \left(\frac{\lambda_0 \theta_1 + (n/\sigma^2)\bar{x}}{\lambda_0 + n/\sigma^2}, \frac{\sigma^2}{n + \lambda_0 \sigma^2} \right).$$

Conjugate Family

We model the rate of production (proteins/cell/hour or car/intersection/minute) by a $\Gamma(\alpha, \beta)$ random variables and take the gamma density as the prior. The arrivals follows a Poisson random variable with parameter λ. Thus the prior density

$$\pi(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

The likelihood for data $\mathbf{x} = (x_1, \ldots, x_n)$,

$$\mathbf{L}(\lambda|\mathbf{x}) = h_\Lambda(\mathbf{x}) \lambda^{T(\mathbf{x})} e^{-n\lambda}$$

where $T(\mathbf{x}) = \sum_{i=1}^n x_i$ is a complete sufficient statistic. The posterior density

$$f_{\Lambda|\mathbf{X}}(\lambda|\mathbf{x}) \propto \mathbf{L}(\lambda|\mathbf{x}) \pi(\lambda|\alpha, \beta) \propto \lambda^{T(\mathbf{x})} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{T(\alpha+\mathbf{x})-1} e^{-(\beta+n)\lambda}$$

which is a density in the gamma family.

Conjugate Family

Thus the gamma family of distributions is a conjugate family for the Poisson family of distributions. In particular If $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$, then

$$(\alpha, \beta) \xrightarrow{\mathbf{x}} (\alpha + T(\mathbf{x}), \beta + n)$$

is the mapping from the parameters in the prior density to the parameters in the posterior for data $\mathbf{x}$.

For example, if we have an prior estimator for a collection (of cells or intersections) with $\alpha = 2.4$ and $\beta = 0.6$ and data $\mathbf{x} = (1, 1, 0, 5, 3, 5, 3, 1, 3, 4)$, then $T(\mathbf{x}) = 26$ and the posterior distribution is $\Gamma(26.4, 10.6)$.

Posterior Predictive Probability

The posterior predictive distribution is the distribution of possible not yet observed value x^* conditioned on the observed values. In symbols,

$$f_{X^*}(x^*) = \int_\Theta f_{X^*|\Psi}(x^*|\theta) f_{\Psi|X}(\theta|\mathbf{x})\ \nu(d\theta)$$

For the case of Bernoulli trials with a $Beta(\alpha, \beta)$ prior, we have posterior $Beta(\alpha + T(\mathbf{x}), \beta + 1 - T(\mathbf{x}))$, $T(\mathbf{x}) = \sum_{i=1}^n x_i$ and

$$f_{X^*}(x^*) = \begin{cases} 0 & \text{with probability } \frac{\alpha + T(\mathbf{x})}{\alpha+\beta+n} \\ 1 & \text{with probability } \frac{\beta+n-T(\mathbf{x})}{\alpha+\beta+n} \end{cases}$$

Posterior Predictive Probability

For a $N(\theta_1, 1/\lambda_0)$ prior and observations $X_1, .., X_n \sim N(\theta, \sigma^2)$, we have posterior

$$N\left(\frac{\lambda_0\theta_1 + (n/\sigma^2)\bar{x}}{\lambda_0 + n/\sigma^2}, \frac{\sigma^2}{n + \lambda_0\sigma^2}\right)$$

The posterior predictive distribution is

$$N\left(\frac{\lambda_0\theta_1 + (n/\sigma^2)\bar{x}}{\lambda_0 + n/\sigma^2}, \frac{\sigma^2}{n + \lambda_0\sigma^2} + \frac{1}{\lambda_0}\right)$$

Posterior Predictive Probability

For $X_1, . , . , X_n \sim Pois(\Lambda)$ with $\Lambda \sim \Gamma(\alpha, \beta)$, we have a $\Gamma(\alpha + T(\mathbf{x}), \beta + n)$ posterior, $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and

$$
\begin{aligned}
f_{X^*}(x^*) &= \int_\Theta f_{X^*|\Lambda}(x^*|\lambda) f_{\Lambda|X}(\lambda|\mathbf{x}) \, d\lambda \\
&= \int_0^\infty \frac{\lambda^{x^*}}{x^*!} e^{-\lambda} \frac{(n+\beta)^{n+T(\mathbf{x})}}{\Gamma(T(\mathbf{x})+\alpha)} \lambda^{T(\mathbf{x})+\alpha-1} e^{-(n+\beta)\lambda} \, d\lambda \\
&= \frac{(n+\beta)^{n+T(\mathbf{x})}}{x^*!\Gamma(T(\mathbf{x})+\alpha)} \int_0^\infty \lambda^{T(\mathbf{x})+\alpha+x^*-1} e^{-(n+\beta+1)\lambda} \, d\lambda \\
&= \frac{(n+\beta)^{n+T(\mathbf{x})}}{x^*!\Gamma(T(\mathbf{x})+\alpha)} \frac{\Gamma(T(\mathbf{x})+\alpha+x^*)}{(n+\beta+1)^{T(\mathbf{x})+\alpha+x^*}} \\
&= \frac{\Gamma(T(\mathbf{x})+\alpha+x^*)}{x^*!\Gamma(T(\mathbf{x})+\alpha)} \left(\frac{n+\beta}{n+\beta+1}\right)^{n+T(\mathbf{x})} \left(\frac{1}{n+\beta+1}\right)^{x^*}
\end{aligned}
$$

and $X^* \sim Negbin\left(T(\mathbf{x}) + \alpha, \left(\frac{1}{n+\beta+1}\right)\right)$