Bayesian Statistics
Introduction

Shaobo Jin

Department of Mathematics

Parametric Statistical Model

Suppose that the vector of observations $x = (x_1, ..., x_n)$ is generated from a probability distribution with density $f(x \mid \theta)$, where θ is the vector of parameters.

- For example, if we further assume the observations are iid, then

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

A parametric statistical model consists of the observation x of a random variable X, distributed according to the density $f(x \mid \theta)$, where the parameter θ belongs to a parameter space Θ of finite dimension.

Likelihood Function

Definition

For an observation x of a random variable X with density $f(x \mid \theta)$, the likelihood function $L(\cdot \mid x) : \Theta \to [0, \infty)$ is defined by $L(\theta \mid x) = f(x \mid \theta)$.

Example

If $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ is a sample of independent random variables, then

$$L(\theta \mid x) = \prod_{i=1}^{n} f_i(x_i \mid \theta),$$

as a function in θ conditional on x.

Likelihood Function: Example

1. If X_1, ..., X_n is a sample of i.i.d. random variables according to $N\left(\theta, \sigma^2\right)$, then

$$L(\theta \mid x) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right].$$

2. If X_1, ..., X_n is a sample of i.i.d. random variables according to $\text{Binomial}\,(k, \theta)$, then

$$L(\theta \mid x) = \prod_{i=1}^{n}\left[\binom{k}{x_i}\theta^{x_i}(1-\theta)^{n-x_i}\right].$$

Likelihood Function: Another Example

Consider the case

- For $i \neq j$, $\begin{bmatrix} X_{i1} & \cdots & X_{in} \end{bmatrix}$ and $\begin{bmatrix} X_{j1} & \cdots & X_{jn} \end{bmatrix}$ are independent and identically distributed.
- For each i, X_{i1}, ..., X_{ip} are not necessarily independent.

Then, the likelihood is

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_{i1}, \cdots, x_{ip} \mid \theta),$$

where $f(x_{i1}, \cdots, x_{ip} \mid \theta)$ is the joint density of $\begin{bmatrix} X_{i1} & \cdots & X_{ip} \end{bmatrix}$.

Inference Principle

In the frequentist context,

1. likelihood principle: the information brought by observation x is entirely contained in the likelihood function $L(\theta \mid x)$.
2. sufficiency principle: two observations x and y factorizing through the same value of a sufficient statistic T as $T(x) = T(y)$ must lead to the same inference on θ.

Bayes Formula

If A and E are two events, then

$$\begin{aligned} \mathrm{P}(A \mid E) &= \frac{\mathrm{P}(E \mid A)\,\mathrm{P}(A)}{\mathrm{P}(E)} \\ &= \frac{\mathrm{P}(E \mid A)\,\mathrm{P}(A)}{\mathrm{P}(E \mid A)\,\mathrm{P}(A) + \mathrm{P}(E \mid A^c)\,\mathrm{P}(A^c)}. \end{aligned}$$

If X and Y are two random variables, then

$$f(y \mid x) = \frac{f(x \mid y)\,f(y)}{f(x)} = \frac{f(x \mid y)\,f(y)}{\int f(x \mid y)\,f(y)\,dy}.$$

Prior and Posterior

A Bayes model consists of a distribution $\pi(\theta)$ on the parameters, and a conditional probability distribution $f(x \mid \theta)$ on the observations.

- The distribution $\pi(\theta)$ is called the prior distribution.
- The unknown parameter θ is a random parameter.

By Bayes formula,

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)} = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)\,d\theta},$$

where the conditional distribution $\pi(\theta \mid x)$ is the posterior distribution and $m(x)$ is the marginal distribution of x.

Update Our Knowledge on θ

The prior often summarizes the prior information about θ.

- From similar experiences, the average number of accidents at a crossing is 1 per 30 days. We assume

$$\pi(\theta) = 30\exp(-30\theta), \quad [\text{day}]^{-1}.$$

Our experiment resulted in an observation x.

- Three accidents have been recorded after monitoring the roundabout for one year. The likelihood is

$$f(X = 3 \mid \theta) = \frac{(365\theta)^3}{3!}\exp(-365\theta).$$

We use the information in x to update our knowledge on θ.

- By Bayes' formula

$$\pi(\theta \mid x) = \frac{f(X = 3 \mid \theta)\,\pi(\theta)}{m(x)}.$$

Distributions

In a Bayesian model, we will have many distributions

- prior distribution: $\pi(\theta)$.
- conditional distribution $X \mid \theta$ (likelihood): $f(x \mid \theta)$.
- joint distribution of (θ, X): $f(x, \theta) = f(x \mid \theta)\,\pi(\theta)$.
- posterior distribution: $\pi(\theta \mid x)$.
- marginal distribution of X: $m(x) = \int f(x \mid \theta)\,\pi(\theta)\,d\theta$.

We most of the time use $\pi(\cdot)$ and $m(\cdot)$ as generic symbols. But in several cases, they are tied to specific functions.

Use Bayes Formula To Obtain Posterior

Example

Find the posterior distribution.

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}(\theta)$, $i = 1, ..., n$. The prior is $\theta \sim \text{Beta}(a_0, b_0)$.
2. Suppose that we have an iid sample $X_i \mid \mu \sim N(\mu, \sigma^2)$, $i = 1, ..., n$, where σ^2 is known. The prior is $\mu \sim N(\mu_0,\ \sigma_0^2)$.
3. Suppose that we have an iid sample $X_i \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$, $i = 1, ..., n$. The priors are $\mu \mid \sigma^2 \sim N(\mu_0, \sigma^2/\lambda_0)$ and $\sigma^2 \sim \text{InvGamma}(a_0, b_0)$, where

$$\pi\left(\sigma^2\right) = \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{1}{(\sigma^2)^{a_0+1}} \exp\left(-\frac{b_0}{\sigma^2}\right).$$

Bayesian Inference Principle

Bayesian Inference Principle

Information on the underlying parameter θ is entirely contained in the posterior distribution $\pi(\theta \mid x)$. That is, all statistical inference are based on the posterior distribution $\pi(\theta \mid x)$.

Some examples are

1. posterior mean: $\mathrm{E}[\theta \mid x]$.
2. posterior mode (MAP): θ that maximizes $\pi(\theta \mid x)$.
3. predictive distribution of a new observation:

$$f(y \mid x) = \int f(y \mid x, \theta)\, \pi(\theta \mid x)\, d\theta.$$

From Univariate to Multivariate Normal

Let $Z \sim N(0,1)$. Then, $X = \sigma Z + \mu \sim N\left(\mu, \sigma^2\right)$, where $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Let $Z = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_p \end{bmatrix}^T$ be a random vector, each $Z_j \sim N(0,1)$, and Z_j is independent of Z_k for any $j \neq k$. Then,

$$X = \Sigma^{1/2} Z + \mu \in \mathbb{R}^p$$

follows a $p-$dimensional multivariate normal distribution, denoted by $X \sim N_p(\mu, \Sigma)$, where $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \Sigma$.

From Univariate to Multivariate Normal: Density

The density function of the random variable $X \sim N\left(\mu, \sigma^2\right)$ with $\sigma > 0$ can be expressed as

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)\right\}.$$

A p-dimensional random variable $X \sim N_p(\mu, \Sigma)$ with $\Sigma > 0$ has the density

$$f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}.$$

Some Useful Properties

1. Linear combination of normal remains normal: Suppose that $X \sim N_p(\mu, \Sigma)$, then $AX + d \sim N_q\left(A\mu + d, A\Sigma A^T\right)$, for every $q \times p$ constant matrix A, and every $p \times 1$ constant vector d.
2. Marginal normal + independence imply joint normal: If X_1 and X_2 are independent and are distributed $N_p(\mu_1, \Sigma_{11})$ and $N_q(\mu_2, \Sigma_{22})$, respectively, then

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \right).$$

3. Conditional distribution: Let $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$. Then the conditional distribution of X_1 given that $X_2 = x_2$, is

$$X_1 \mid X_2 \sim N\left\{ \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right\}.$$

Multivariate Normal In Bayesian Statistics

Example

Suppose that $X \mid \theta \sim N_p\left(C\theta, \Sigma\right)$, where $C_{p\times q}$ and $\Sigma > 0$ are known. The prior is $N_q\left(\mu_0,\, \Lambda_0^{-1}\right)$. Find the posterior of θ.

We can in fact use the property of the conditional distribution of a multivariate normal distribution to simplify the steps.

Result

If we know $X_1 \mid X_2 \sim N_p\left(CX_2, \Sigma\right)$ and $X_2 \sim N_q\left(m, \Omega\right)$, then

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} Cm \\ m \end{bmatrix}, \begin{bmatrix} \Sigma + C\Omega C^T & C\Omega \\ \Omega C^T & \Omega \end{bmatrix}\right).$$