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 $f\left(x \mid \theta\right)$, the likelihood function $L\left(\cdot \mid x\right) : \Theta \to [0, \infty)$ is defined by $L\left(\theta \mid x\right) = f\left(x \mid \theta\right)$.

Example

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

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

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

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

2. If X_1, ..., X_n is a sample of i.i.d. random variables according to Binomial (k, θ), then

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

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

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

where $f\left(x_{i1}, \cdots, x_{ip} \mid \theta\right)$ 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 $\pi\left(\theta\right)$ on the parameters, and a conditional probability distribution $f\left(x \mid \theta\right)$ on the observations.

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

By Bayes formula,

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

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

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

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

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

Bayesian Statistics
Prior

Shaobo Jin

Department of Mathematics

Prior Distribution

The main difference between a frequentist model and a Bayesian model is that the parameter of the data generating distribution is random and follows a known distribution (prior distribution). The parameters in a prior distribution are called the hyperparameters.

1. A subjective prior incorporates our prior knowledge.
2. An objective prior fulfills some desired (theoretical) properties.

It is in general very difficult to specify an exact prior distribution. Most critiques of Bayesian methods is specifying a prior distribution.

Subjective Prior: Expert Advise

Example

Suppose that we are interested in the effectiveness $\theta \in [0, 1]$ of a vaccine.

- An expert expects a 80% decrease in the number of disease cases among the group of vaccinated people compared to non-vaccinated group of people.
- Suppose that we would like to use a $\text{Beta}\,(a, b)$ prior.
- The hyperparameters can be set such that the expectation of the beta distribution $\frac{a}{a+b}$ is close to 80%.

Subjective Prior: Previous Experiences

Example

Suppose that we want to predict the number of sold cups of coffee during midsommar celebration.

- Suppose that the sales records from previous years show that the number ranges between 600 and 800 cups.
- We can choose a prior distribution such that the majority of mass is close/within such range.

Mixture Prior Distribution: Example

Example

Suppose that we are interested in the temperature θ at the midsommar celebration.

- One expert guesses that the temperature is around 22°C, and another expert guesses 10°C.
- One example is to specify the temperate as

$$wN\left(22,\sigma_1^2\right)+(1-w)\,N\left(10,\sigma_2^2\right).$$

Conjugate Prior

Definition

Let $\mathcal{F}$ be a family of probability distributions on Θ. If $\pi\left(\cdot\right) \in \mathcal{F}$ and $\pi\left(\cdot \mid x\right) \in \mathcal{F}$ for every x, then the family of distributions $\mathcal{F}$ is conjugate. The prior distribution that is an element in a conjugate family is called a conjugate prior.

The main benefit of a conjugate prior is tractability, that is, we only need to update the hyperparameters without changing the family of distributions. It makes Bayesian computation much easier.

Conjugate Prior: Example

Example

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

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

Find Conjugate Prior

The likelihood $f\left(x \mid \theta\right)$ entirely determines the class of conjugate priors.

Example

Find the conjugate prior.

1. Suppose that we have an iid sample $X_i \mid \theta \sim \text{Poisson}\left(\theta\right)$.
2. Suppose that we have an iid sample

$$X_i \mid \theta \quad \sim \quad \text{Multinomial}\left(m, \theta_1, ..., \theta_k\right).$$

Exponential Family

Definition

A class of probability distributions $\mathcal{P} = \{\mathrm{P}_\theta : \ \theta \in \Theta\}$ is called an exponential family, if there exist a number $k \in \mathbb{N}$, real-valued functions A, ζ_1, .., ζ_k on Θ, real-valued statistics T_1, ..., T_k, and a function h on the sample space $\mathcal{X}$ such that

$$f\left(x \mid \theta\right) = A\left(\theta\right) \exp\left\{\sum_{j=1}^{k} \zeta_j\left(\theta\right) T_j\left(x\right)\right\} h\left(x\right),$$

where $A\left(\theta\right) > 0$ depends only on θ and $h\left(x\right) \geq 0$ depends only on x. We often denote the real valued functions by

$$\begin{aligned} \zeta\left(\theta\right) &= \begin{pmatrix} \zeta_1 & \cdots & \zeta_k \end{pmatrix}^T, \\ T\left(x\right) &= \begin{pmatrix} T_1 & \cdots & T_k \end{pmatrix}^T. \end{aligned}$$

Exponential Family: Example

Example (Normal distribution)

Normal distribution with $\theta = \left(\mu, \sigma^2\right)$:

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

Example (Binomial distribution)

Binomial distribution:

$$\mathrm{P}\left(X = x \mid \theta\right) = \exp\left\{x \log\left(\theta\right) + \left(n - x\right)\log\left(1 - \theta\right)\right\} \binom{n}{x}.$$

Exponential Family: Counterexample

1. Exponential distribution:

$$\begin{aligned} f(x \mid \theta) &= \theta \exp\{-\theta x\}, \quad x \geq 0 \\ &= \theta \exp\{-\theta x\} 1(x \geq 0), \end{aligned}$$

   where $1(\cdot)$ is the indicator function.

2. Shifted exponential distribution with $\theta = (\lambda, \mu)$:

$$\begin{aligned} f(x \mid \theta) &= \lambda \exp\{-\lambda(x-\mu)\}, \quad x \geq \mu \\ &= \lambda \exp\{\lambda\mu\} \exp\{-\lambda x\} 1(x \geq \mu). \end{aligned}$$

Natural Parameter

We can parameterize the probability function as

$$f\left(x \mid \zeta\right) = C\left(\zeta\right)\exp\left\{\sum_{j=1}^{k}\zeta_j T_j\left(x\right)\right\} h\left(x\right),$$

where ζ is called the natural parameter.

Example (Binomial distributin)

For $\theta \in (0,1)$,

$$f\left(x \mid \theta\right) = \left(1-\theta\right)^n \exp\left\{x\log\left(\frac{\theta}{1-\theta}\right)\right\}\binom{n}{x}.$$

Define $\zeta = \log\left(\frac{\theta}{1-\theta}\right) \in \mathbb{R}$. Then,

$$f\left(x \mid \zeta\right) = \left(1-\frac{\exp\left(\zeta\right)}{1+\exp\left(\zeta\right)}\right)^n \exp\left\{x\zeta\right\}\binom{n}{x}.$$

Conjugate Prior for Exponential Family

Theorem

Suppose that

$$f(x \mid \zeta) = \exp\left\{\sum_{j=1}^{k} \zeta_j T_j(x) + \log C(\zeta)\right\} h(x).$$

Then the conjugate family for ζ *is given by*

$$\pi(\zeta) = K(\mu_0, \lambda_0) \exp\left\{\zeta^T \mu_0 + \lambda_0 \log C(\zeta)\right\},$$

where μ *and* λ *are hyperparameters. The posterior satisfies*

$$\pi(\zeta \mid x) \propto \exp\left\{\zeta^T [\mu_0 + T(x)] + (\lambda_0 + 1) \log C(\zeta)\right\}.$$

Conjugate Prior: Example

Example

Using the exponential family for the following examples.

1. Let X_1, .., X_n be an iid from $N\left(\theta, \sigma^2\right)$, where σ^2 is known. Show that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$ is conjugate.
2. Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}\left(\theta\right)$. Show that $\theta \sim \text{Beta}\left(a, b\right)$ is conjugate.
3. Suppose that $\{Y_i\}_{i=1}^n$ are independent observations such that

$$\mathrm{P}\left(Y_i = 1 \mid X_i = x_i\right) \quad = \quad \frac{\exp\left\{y_i\left(x_i^T\theta\right)\right\}}{1+\exp\left(x_i^T\theta\right)}.$$

   Find the conjugate prior for θ.

No Prior Information

When no prior information is available, we still need to specify a prior in order to use Bayesian modelling.

Definition

The Laplace prior is $\pi\left(\theta\right)$ is a constant over Θ.

Disambiguation: The name Laplace prior is often referred to as that θ follows a Laplace distribution

$$\pi\left(\theta\right) = \frac{1}{2b_0}\exp\left(-\frac{|\theta - a_0|}{b_0}\right), \quad -\infty < \theta < \infty.$$

The prior in the above definition is often referred to as the flat prior, uniform prior, among others.

Uniform Prior as Non-Informative Prior

Intuitively speaking, a constant $\pi(\theta)$ means that we treat all θ equally.

- The posterior depends only on the likelihood.

For a distribution P with density p, its entropy is

$$S(P) = -\mathrm{E}[\log p].$$

The entropy is often called the Shannon entropy if the random variable is discrete and the differential entropy if the random variable is continuous.

Example

Find the entropy of the following distributions.

1. $X \sim N(0, \sigma^2)$.
2. X is uniform on the finite discrete set $\{1, 2, ..., n\}$.

Uniform Distribution Maximizes Entropy

The entropy of a random variable measures its uncertainty.

- If a random variable puts majority of probability mass on one value, then the uncertainty is small.
- If the possible values of a random variable are equally alike, then the uncertainty is large.

Example

1. Suppose that X is a discrete random variable with a finite sample space $\{1, 2, ..., n\}$. Show that the discrete uniform distribution maximizes the Shannon entropy.
2. Suppose that X is a continuous random variable with a closed sample space $[a, b]$. Show that the continuous uniform distribution maximizes the differential entropy.

Improper Prior

The uniform prior is proportional to a density of a probability measure if the parameter space Θ is bounded.

However, in many cases, the prior is not a probability measure. Instead it yields

$$\int_{\Theta} \pi(\theta)\, d\theta = \infty.$$

Such prior distribution is said to be an improper prior.

- The uniform prior is an improper prior if Θ is not bounded.

But as long as the posterior distribution is well defined, the Bayesian methods still apply.

Improper Posterior

One risk of using improper prior is that the posterior can be undefined.

Example

Let $X \sim \text{Binomial}\left(n, \theta\right)$ and $\pi\left(\theta\right) \propto \frac{1}{\theta(1-\theta)}$. The posterior satisfies

$$\begin{aligned} \pi\left(\theta \mid x\right) &\propto \theta^{x}\left(1-\theta\right)^{n-x} \frac{1}{\theta\left(1-\theta\right)} \\ &= \theta^{x-1}\left(1-\theta\right)^{n-x-1}, \end{aligned}$$

which is not defined for $x = 0$ or $x = n$.

In order to have a well-defined posterior, we need

$$\int f\left(x \mid \theta\right) \pi\left(\theta\right) d\theta < \infty.$$

But this may not be an easy task to check.

Marginalization Paradox

Since the improper prior is not a probability density, the posterior, even exists, may not follow the rules of probability. One example is the marginalization paradox.

- Consider a model $f(x \mid \alpha, \beta)$ and a prior $\pi(\alpha, \beta)$. Suppose that the marginal posterior $\pi(\alpha \mid x)$ satisfies

$$\pi(\alpha \mid x) = \pi(\alpha \mid z(x))$$

  for some function $z(x)$.
- Suppose that $f(z \mid \alpha, \beta) = f(z \mid \alpha)$, that is, does not depend on β.
- If $\pi(\alpha, \beta)$ is a proper prior, we can recover $\pi(\alpha \mid x)$ from $f(z \mid \alpha)$ and some $\pi(\alpha)$ as $\pi(\alpha \mid x) \propto f(z \mid \alpha)\pi(\alpha)$.
- However, if $\pi(\alpha, \beta)$ is not a proper prior, it can happen that $f(z \mid \alpha)\pi(\alpha)$ is not proportional to $\pi(\alpha \mid x)$ for any $\pi(\alpha)$.

Marginalization Paradox: Example

Example

Let X_1, ..., X_n be independent exponential random variables. The first m have mean η^{-1} and the rest have mean $(c\eta)^{-1}$, where $c \neq 1$ is a known constant and $m \in \{1, ..., n-1\}$.

- We consider the improper prior $\pi(\eta) = 1$ such that $\pi(\eta, m) = \pi(\eta)\pi(m) = \pi(m)$.
- The marginal posterior distribution satisfies

$$\pi(m \mid x) \quad \propto \quad \frac{c^{n-m}\pi(m)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n+1}},$$

  where $z_i = x_i/x_1$. Hence, the marginal posterior depends only on $z = (z_2, ..., z_n)$, since $z_1 = 1$.

Marginalization Paradox: Example

Example

$$\pi\left(m \mid x\right) \quad \propto \quad \frac{c^{n-m}\pi\left(m\right)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n+1}}.$$

- The density of z is

$$f\left(z \mid \eta, m\right) \quad = \quad \frac{c^{n-m}\Gamma\left(n\right)}{\left(\sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i\right)^{n}} \equiv f\left(z \mid m\right),$$

  which only depends on m, not η.
- However, it is not possible to find a $\pi^*\left(m\right)$ such that

$$\pi\left(m \mid x\right) \quad \propto \quad f\left(z \mid m\right)\pi^*\left(m\right).$$

Invariance?

Another issue of the uniform prior is that it is not invariant against reparametrization.

- Suppose that we choose the uniform prior for $\theta \in \Theta$.
- Now we reparameterize to $\eta = \eta\left(\theta\right)$, which is one-to-one, such that $\theta = h\left(\eta\right)$. Then,

$$\pi_{\eta}\left(\eta\right) \quad = \quad \pi_{\theta}\left(h\left(\eta\right)\right)\left|\det\left[\frac{\partial h\left(\eta\right)}{\partial \eta^{T}}\right]\right|,$$

  which is not a constant.

A constant prior on θ does not always yield a constant prior on $\eta\left(\theta\right)$, even though η is a strictly monotone transformation.

Invariance: Example

Example (Binomial distributin)

Suppose that $X \mid \theta \sim \text{Binomial}(n, \theta)$. We have no information regarding θ. Hence we let $\theta \sim \text{Uniform}(0,1)$.

- Consider the odds ratio $\zeta = \frac{\theta}{1-\theta}$.
- By change of variables,

$$\begin{aligned}\pi_\zeta(\zeta) &= \pi_\theta(h(\zeta)) \left|\det\left[\frac{\partial h(\eta)}{\partial \eta^T}\right]\right| = \left|\frac{\partial}{\partial \zeta}\frac{\zeta}{1+\zeta}\right| \\ &= \frac{1}{(1+\zeta)^2}.\end{aligned}$$

- Further, $\theta \sim \text{Uniform}(0,1)$ is the same as $\theta \sim \text{Beta}(1,1)$. The prior is conjugate. But the resulting prior for ζ is not.

Invariance Under Monotone Transformation

Suppose that a procedure of finding prior yields the prior density $\pi_\theta(\theta)$ for θ.

- Let h be a smooth and monotone transformation. By the change of variables $\eta = \eta(\theta)$ and $\theta = h(\eta)$, the density of η induced from $\pi_\theta(\theta)$ is

$$\pi_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| .$$

- If we use the same procedure of finding prior as we used for θ, it should yield the prior density $\pi_\eta(\eta)$ for η.

Invariance means that such two densities should be the same:

$$\pi_\theta(h(\eta)) \left| \frac{dh(\eta)}{d\eta} \right| = \pi_\eta(\eta) .$$

Motivation: Use of Fisher Information

Suppose that P and Q are two probability measures with densities p and q, respectively.

- The Kullback-Leibler divergence is

$$\mathrm{KL}\left(P, Q\right) \quad = \quad \int \log\left[\frac{p\left(x\right)}{q\left(x\right)}\right] p\left(x\right) dx.$$

- We consider the symmetric metric

$$\mathrm{KL}\left(P_{\theta}, P_{\theta'}\right) + \mathrm{KL}\left(P_{\theta'}, P_{\theta}\right).$$

- If we change the parametrization such that $\theta = h\left(\eta\right)$ using Fisher information, then parametrization leaves the distance between distributions approximately unchanged:

$$\mathrm{KL}\left(P_{\theta}, P_{\theta'}\right) + \mathrm{KL}\left(P_{\theta'}, P_{\theta}\right) \quad = \quad \mathrm{KL}\left(P_{\eta}, P_{\eta'}\right) + \mathrm{KL}\left(P_{\eta'}, P_{\eta}\right).$$

Jeffreys Prior

Definition

Consider a statistical model $f\left(x \mid \theta\right)$ with Fisher information matrix $\mathcal{I}\left(\theta\right)$. The Jeffreys prior is

$$\pi\left(\theta\right) \quad \propto \quad \left[\det\left(\mathcal{I}\left(\theta\right)\right)\right]^{1/2}.$$

The Jeffreys prior is invariant to reparametrization under smooth monotone transformation, because we can show

$$\pi_{\theta}\left(h\left(\eta\right)\right)\left|\frac{dh\left(\eta\right)}{d\eta}\right| \quad = \quad \pi_{\eta}\left(\eta\right).$$

Jeffreys Prior: Example

Example

Find the Jeffreys prior for θ.

1. Suppose that $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. Show also that the Jeffreys prior is invariant to the transformation $\eta = \frac{\theta}{1-\theta}$.
2. Suppose that $X_i \mid \theta \sim N\left(\theta, 1\right)$, $i = 1, ..., n$.
3. Suppose that $X_i \mid \theta$ belongs to a location family with density $f\left(x_i - \theta\right)$, where $f\left(x\right)$ is a density function.
4. Suppose that $X_i \mid \theta$ belongs to a scale family with density $\theta^{-1} f\left(\theta^{-1} x_i\right)$, where $f\left(x\right)$ is a density function and $\theta \in \mathbb{R}_+$.

Jeffreys Prior is Non-Informative

The Jeffreys prior is derived in order to achieve invariance. It turns out that it is also non-informative.

- Under the Jeffreys prior, the posterior can be approximated by

$$\begin{aligned} \pi(\theta \mid x) &\propto \pi(\theta) f(x \mid \theta) \\ &\approx [\det(\mathcal{I}(\theta))]^{1/2} \exp\left(-\frac{1}{2}\left(\theta-\hat{\theta}\right)^T \mathcal{I}(\theta)\left(\theta-\hat{\theta}\right)\right), \end{aligned}$$

  that is, $\theta \mid x \approx N\left(\hat{\theta}, \mathcal{I}^{-1}(\theta)\right)$.

- The frequentist approach yields $\hat{\theta}-\theta \approx N\left(0, \mathcal{I}^{-1}(\theta)\right)$.

Inference using the Jeffreys prior coincides approximately with the inference from the likelihood function.

Independent Jeffreys Prior

Example

Suppose that $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Find the Jeffreys prior for $\theta = \left(\mu, \sigma^2\right)$.

When we have multiple parameters, it is also common to use the independent Jeffreys prior.

- Obtain the Jeffeys prior for each parameter separately by fixing the others.
- Multiple the single parameter Jeffreys prior together.

Example

Suppose that $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. Find the independent Jeffreys prior for $\theta = \left(\mu, \sigma^2\right)$.

A Cautious Note

The Jeffreys prior do not necessarily perform satisfactorily for all inferential purposes.

Example

Suppose that we observe one observation $X \mid \theta \sim N_p(\theta, I)$.

- The Jeffreys prior is the uniform prior and the posterior is $\theta \mid x \sim N_p(x, I)$.
- Suppose that we are interested in the parameter $\eta = \theta^T \theta$. The posterior distribution of η is noncentral χ^2 with p degrees of freedom. The posterior expected value is $x^T x + p$.
- If we consider a quadratic loss, the loss of another estimator $x^T x - p$ is no greater than the loss of $x^T x + p$ for all θ.
- This means that for any θ, we can always find an estimator that is better than the estimator using the Jeffreys prior.

Reference Prior

Consider the Kullback-Leibler divergence

$$\mathrm{KL}\left(\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right) = \int \pi\left(\theta \mid x\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta \geq 0.$$

A large KL means that a lot information has come from the data.

The expected KL under the marginal of x is then

$$\begin{aligned} \mathrm{E}\left[\mathrm{KL}\left(\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right)\right] &= \int m\left(x\right)\left[\int \pi\left(\theta \mid x\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta\right] dx \\ &= \int\int f\left(x, \theta\right) \log\left(\frac{\pi\left(\theta \mid x\right)}{\pi\left(\theta\right)}\right) d\theta dx \\ &= \int\int f\left(x, \theta\right) \log\left(\frac{f\left(x, \theta\right)}{\pi\left(\theta\right) m\left(x\right)}\right) d\theta dx, \end{aligned}$$

where $m\left(x\right)$ is the marginal density of x.

Mutual Information

In probability theory, the mutual information of two random variables X and Y is defined as

$$\mathrm{MI}\left(X,Y\right)=\int\int f\left(x,y\right)\log\left(\frac{f\left(x,y\right)}{f\left(x\right)f\left(y\right)}\right)dxdy \quad \geq \quad 0.$$

It is a measure to quantity the information in $f\left(x,y\right)$ instead of $f\left(x\right)f\left(y\right)$.

- The expected KL in the previous slide

$$\mathrm{E}\left[\mathrm{KL}\left(\pi\left(\theta\mid x\right),\pi\left(\theta\right)\right)\right] \quad = \quad \int\int f\left(x,\theta\right)\log\left(\frac{f\left(x,\theta\right)}{\pi\left(\theta\right)m\left(x\right)}\right)d\theta dx$$

  is the mutual information of X and θ.

The reference prior aims to maximize the mutual information of the prior and posterior.

Reference Prior and Entropy

Result

Let $p(x)$ be the density of a distribution P. Then,

$$\mathrm{MI}(X,\theta) = S(\pi(\theta)) - \int m(x) S(\pi(\theta \mid x)) dx,$$

where

$$S(P) = -\mathrm{E}[\log p(X)]$$

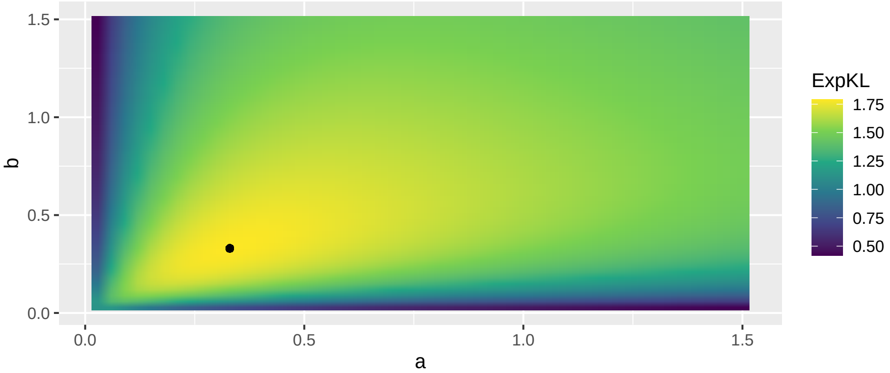
is the entropy.

Thus, a reference prior that generates a large mutual information corresponds to a prior with large entropy and a posterior with low expected entropy.

Reference Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and we consider the class of conjugate priors $\theta \sim \text{Beta}\,(a, b)$. Find the expected KL.



Explicit Form of Reference Prior

Suppose that we can replicate the experiment independently k times. Each time we observe a data set of sample size n. Denote all realizations by $x = \left(x^{(1)}, ..., x^{(k)}\right)$.

Let $\pi^* (\theta)$ be a continuous and strictly positive function such that the posterior $\pi^* (\theta \mid x)$ is proper and asymptotically consistent. For any interior point θ_0 of Θ, define

$$\begin{aligned} p_k (\theta) &= \exp \left\{ \int f (x \mid \theta) \log \pi^* (\theta \mid x) \, dx \right\}, \\ p (\theta) &= \lim_{k \to \infty} \frac{p_k (\theta)}{p_k (\theta_0)}, \end{aligned}$$

Suppose $p_k (\theta)$ is continuous for all k. Then, under some extra assumptions on the ratio $p_k (\theta) / p_k (\theta_0)$ and on $p (\theta)$, $p (\theta)$ is a reference prior.

Approximate Reference Prior

Find the reference prior is not an easy task, since the integrals can be difficult to evaluate.

Algorithm 1: Approximate reference prior

```
Choose an arbitrary continuous and positive function $\pi^*(\theta)$, e.g., $\pi^*(\theta) = 1$ ;
for any $\theta$ of interest including a $\theta_0$ do
    for $j$ from 1 to $m$ do
        Simulate independently $\left\{x_j^{(1)}, ..., x_j^{(k)}\right\}$ from $f(x \mid \theta)$ ;
        Compute the integral $c_j = \int_\Theta \left[\prod_{i=1}^k f\left(x_j^{(i)} \mid \theta\right)\right] \pi^*(\theta)\, d\theta$ analytically or approximate numerically ;
        Evaluate $r_j(\theta) = \log\left\{\left[\prod_{i=1}^k f\left(x_j^{(i)} \mid \theta\right)\right] \pi^*(\theta) / c_j\right\}$ ;
    end
    Compute $p_k(\theta) = \exp\left\{m^{-1} \sum_{j=1}^m r_j(\theta)\right\}$ ;
    Let $\pi(\theta) \propto p_k(\theta) / p_k(\theta_0)$ ;
end
```

Reference Prior and Jeffreys Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}\,(n, \theta)$. Approximate the reference prior.

In fact, if the distribution of MLE $\sqrt{n}\left(\hat{\theta} - \theta\right)$ can be approximated by $N\left(\theta,\, \mathcal{I}^{-1}\left(\theta\right)\right)$, and the posterior distribution of $\sqrt{n}\left(\theta - \hat{\theta}\right)$ can be approximately $N\left(0, \mathcal{I}^{-1}\left(\theta\right)\right)$, then, the reference prior and the joint Jeffreys prior are asymptotically equivalent.

Reference Prior With Presence of Nuisance Parameter

Suppose that $x \mid \theta \sim f(x \mid \theta_1, \theta_2)$ and $\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is the nuisance parameter. The reference prior is obtained as follows.

- First, treating θ_1 as fixed. Use the Jeffreys prior associated with $f(x \mid \theta_2)$ as $\pi(\theta_2 \mid \theta_1)$.
- Then, derive the marginal distribution

$$f(x \mid \theta_1) = \int f(x \mid \theta_1, \theta_2) \pi(\theta_2 \mid \theta_1) d\theta_2.$$

Compute the Jeffreys prior $\pi(\theta_1)$ associated with $f(x \mid \theta_1)$.

Neyman-Scott Problem: Example

Consider the Neyman-Scott problem, where $X_{ij} \mid \theta \sim N\left(\mu_{ij}, \sigma^2\right)$, $i = 1, ..., n$ and $j = 1, 2$. We are interested in σ and μ_{ij}'s are nuisance parameters.

- The usual Jeffreys prior is $\pi\left(\theta\right) \propto \sigma^{-n-1}$. The posterior mean of σ^2 is

$$\mathrm{E}\left[\sigma^2 \mid x\right] = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{4n-4} \xrightarrow{P} \frac{\sigma^2}{2} \neq \sigma^2,$$

  where $\xrightarrow{P}$ means convergence in probability.

- The reference prior is $\pi\left(\theta\right) \propto \sigma^{-1}$. The posterior mean of σ^2 is

$$\mathrm{E}\left[\sigma^2 \mid x\right] = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{2n-4} \xrightarrow{P} \sigma^2.$$

Berger-Bernardo Method

The idea of deriving the prior conditioning on a subset of parameter can be applied to a general setting with more than two sets of parameters. The resulting method is the Berger-Bernado method.

Suppose the $p \times 1$ vector θ is partitioned into m groups, denoted by θ_1,..., θ_m. The reference prior is obtained in a similar manner to

$$\pi(\theta) \quad \propto \quad \pi(\theta_m \mid \theta_1, ..., \theta_{m-1})\, \pi(\theta_{m-1} \mid \theta_1, ..., \theta_{m-2}) \cdots \pi(\theta_2 \mid \theta_1)\, \pi(\theta_1).$$

Berger-Bernardo Method: Algorithm

```
Algorithm 2: Berger-Bernardo method
Initiate some $\pi_m\left(\theta_m \mid \theta_1, ..., \theta_{m-1}\right)$, e.g., Jeffreys prior ;
for $j$ in $m-1,\ m-2,\ ...,\ 1$ do
    Obtain the marginal distribution
        $f\left(x \mid \theta_1, ..., \theta_j\right) = \int f\left(x \mid \theta\right) \pi_{j+1}\left(\theta_{j+1}, ..., \theta_m \mid \theta_1, ..., \theta_j\right) d\left(\theta_{j+1}, ..., \theta_m\right).$
    Determine the reference prior $h_j\left(\theta_j \mid \theta_1, ..., \theta_{j-1}\right)$ related to the
     model $f\left(x \mid \theta_1, ..., \theta_j\right)$, where $\theta_1, ..., \theta_{j-1}$ is treated as fixed ;
    Compute $\pi_j\left(\theta_j, ..., \theta_m \mid \theta_1, ..., \theta_{j-1}\right)$ by
        $\pi_j\left(\theta_j, ..., \theta_m \mid \theta_1, ..., \theta_{j-1}\right) \propto \pi_{j+1}\left(\theta_{j+1}, ..., \theta_m \mid \theta_1, ..., \theta_j\right) h_j\left(\theta_j \mid \theta_1, ..., \theta_{j-1}\right).$
end
Obtain the reference prior $\pi\left(\theta\right) = \pi_1\left(\theta_1, ..., \theta_m\right)$ ;
```

Berger-Bernardo Method: Example

Example

Consider $X \mid \theta \sim \text{Multinomial}\left(n, \theta_1, ..., \theta_4\right)$. The likelihood is

$$f\left(x \mid \theta_1, \theta_2, \theta_3\right) = \frac{n!}{x_1! x_2! x_3! x_4!} \theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3} \theta_4^{x_4},$$

where $\theta_4 = 1 - \sum_{i=1}^{3} \theta_i$. Find the reference prior of $\theta = (\theta_1, \theta_2, \theta_3)$, where $m = 3$.

Influence of Prior

The assessment of the influence of the prior is called sensitivity analysis. In general, the prior can have a big impact for small sample sizes.

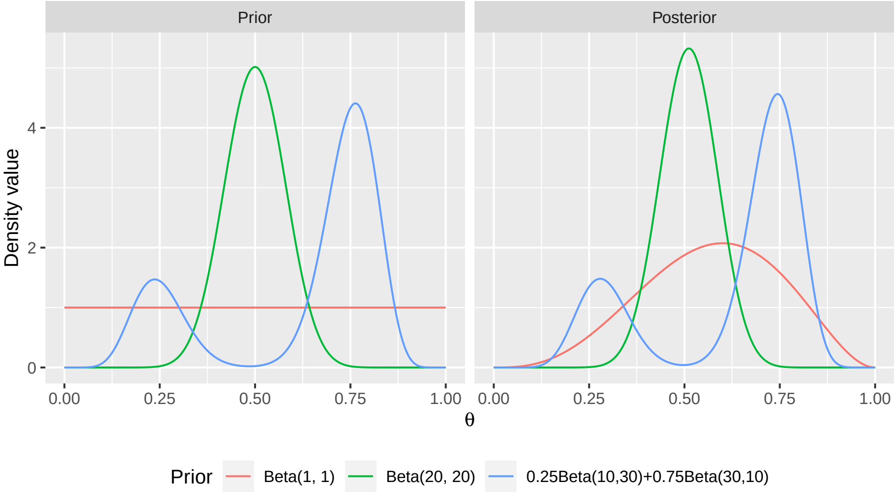
- But it becomes less important as the sample size increases. Most priors will lead to similar inference that is equivalent to the one based only on the likelihood.

Example

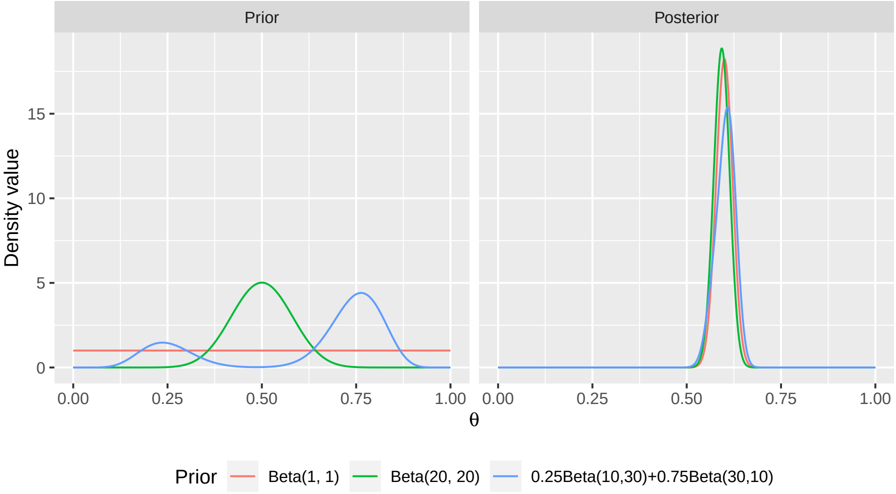
Suppose that we have an iid sample $X_i \mid \theta \sim \text{Bernoulli}\left(\theta\right)$.The conjugate prior $\theta \sim \text{Beta}\left(a_0, b_0\right)$ yields the posterior

$$\text{Beta}\left(a_0 + \sum_{i=1}^{n} x_i, b_0 + n - \sum_{i=1}^{n} x_i\right).$$

Small Sample Size



Large Sample Size



Hierarchical Prior Distribution

We can apply a hierarchical prior, applying a prior on the prior.

- Suppose that $\pi_1(\theta \mid \lambda)$ is a conjugate prior for $f(x \mid \theta)$, where λ is the hyperparameter.
- Instead of specifying the value of λ, we let

$$\lambda \sim \pi_2(\lambda)\,,\ \theta \mid \lambda \sim \pi_1(\theta \mid \lambda)\,,\ x \mid \theta \sim f(x \mid \theta)\,.$$

- For example, if $X \mid z \sim N\left(\mu, z\sigma^2\right)$ and z is inverse gamma, then X follows a t distribution.
- A t distribution prior with low degrees of freedom (e.g., 3) is a popular choice.

Different Priors in Practice

We have introduced different ways of constructing the prior, e.g., conjugate prior, uniform prior, Jeffreys prior, and reference prior. Depending on how much information the priors contain, we can roughly partition the prior into the following groups according to their level of informative relative to the likelihood:

1. noninformative flat prior,
2. super-vague but proper prior, e.g., a prior with a massive variance such as $1,000,000$,
3. very weakly informative prior, e.g., a prior with a sizable variance such as 10,
4. weakly informative prior, e.g., a prior with variance 1,
5. informative prior.

The first two groups are generally not recommended.

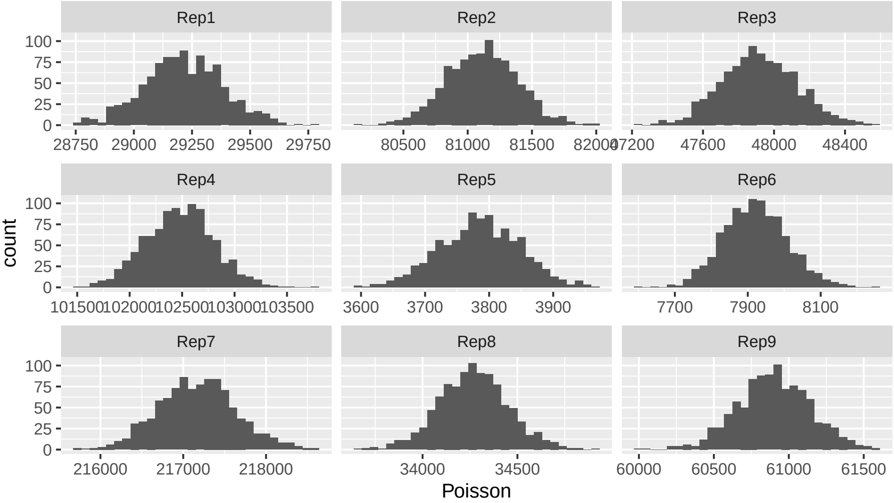
Prior Predictive Check

The prior predictive check is a way to assess whether your prior is appropriate.

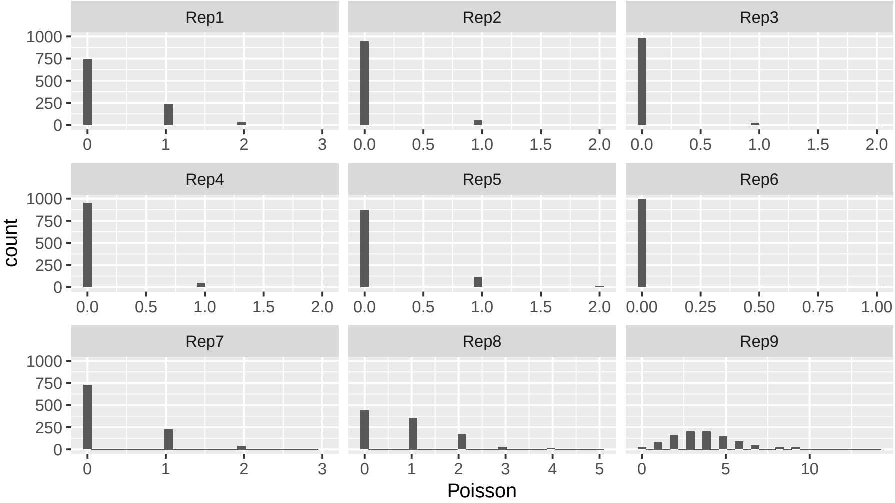
Algorithm 3: Prior predictive check

```
for $j$ in 1, 2, ..., $m$ do
    Simulate $\theta_{\text{sim}} \sim \pi(\theta)$ ;
    Simulate $x_{\text{sim}} \sim f(x \mid \theta_{\text{sim}})$ of sample size $n$ ;
end
Visualize each data set or investigate the summary statistics to judge whether the simulated data are plausible to avoid super bad priors.
```

Prior 1


Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
Rep9
count
100
75
50
25
0
28750 29000 29250 29500 29750
80500 81000 81500 82000
47200 47600 48000 48400
101500 102000 102500 103000 103500
3600 3700 3800 3900
7700 7900 8100
216000 217000 218000
34000 34500
60000 60500 61000 61500
Poisson

Prior 2


Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
Rep9
count
1000
750
500
250
0
0
1
2
3
0.0
0.5
1.0
1.5
2.0
0.00
0.25
0.50
0.75
1.00
4
5
10
Poisson

Bayesian Statistics
Bayesian Estimation

Shaobo Jin

Department of Mathematics

Maximum a Posteriori Estimator

In a Bayes model, the parameter θ is a random variable with known distribution π.

- Finding the true parameter makes no sense in a Bayes model.

Data that we observe are generated in a hierarchical manner:

$$\theta \sim \pi(\theta), \qquad X \mid \theta \sim f(x \mid \theta).$$

Given the data x, we can make inference for the “current” data generating process.

Definition

The maximum a posteriori (MAP) estimator is the mode of the posterior $\pi(\theta \mid x)$ as

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \pi(\theta \mid x).$$

MAP: Example

Note that $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ and $m(x)$ does not include any θ.

- The MAP estimator only requires the kernel of $f(x \mid \theta)\pi(\theta)$.
- We can skip the integration step to obtain $m(x)$.

Example

Let X_1, ..., X_n be iid $N\left(0, \sigma^2\right)$. The parameter of interest is $\theta = \sigma^{-2}$. We assume that the prior of θ is $\text{Gamma}(a, b)$. Find the MAP estimator.

MAP and 0-1 Loss

For an estimator d, consider the loss function

$$L\left(\theta, d\right) = 1\left(\|\theta - d\| > \epsilon\right),$$

where $\|\theta - d\| = \sqrt{\left(\theta - d\right)^T\left(\theta - d\right)}$. Consider the expected value of $L\left(\theta, d\right)$ under the posterior distribution:

$$\begin{aligned} \mathrm{E}\left[L\left(\theta, d\right) \mid x\right] &= \int 1\left(\|\theta - d\| > \epsilon\right)\pi\left(\theta \mid x\right) d\theta \\ &= 1 - \mathrm{P}\left(\|\theta - d\| \leq \epsilon \mid x\right). \end{aligned}$$

To minimize the expected loss, we want $\mathrm{P}\left(\|\theta - d\| \leq \epsilon \mid x\right)$ to be as large as possible, that is, the distribution of $\theta \mid x$ is concentrated around d.

Nuisance Parameter

Suppose that data are generated from $f\left(x \mid \theta, \tau\right)$, where θ is the parameter of interest and τ is the nuisance parameter.

- The frequentist approach will find a sufficient statistic $T\left(x\right)$ for τ and make inference for θ using the conditional distribution of $X \mid T\left(X\right)$.
- Alternatively, inference is based on the profile likelihood

$$L\left(\theta\right) = \max_{\tau} f\left(x \mid \theta, \tau\right) = f\left(x \mid \theta, \hat{\tau}\left(\theta\right)\right),$$

  where $\hat{\tau}\left(\theta\right)$ maximizes $f\left(x \mid \theta, \tau\right)$ for fixed θ.

An advantage of the Bayes approach is that we can simply integrate out the nuisance parameter τ and make inference from the marginal posterior $\pi\left(\theta \mid x\right)$, instead of the joint posterior $\pi\left(\theta, \tau \mid x\right)$.

Neyman-Scott Problem: Example

Consider the Neyman-Scott problem, where $X_{ij} \mid \theta \sim N\left(\mu_{ij}, \sigma^2\right)$, $i = 1, ..., n$ and $j = 1, 2$. We are interested in σ^2, and μ_{ij}'s are nuisance parameters.

- The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{4n} \xrightarrow{P} \frac{\sigma^2}{2} \neq \sigma^2.$$

- Consider the reference prior $\pi\left(\theta\right) \propto \sigma^{-1}$. The MAP of $\pi\left(\sigma^2 \mid x\right)$ is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(x_{i1} - x_{i2}\right)^2}{2n+4} \xrightarrow{P} \sigma^2.$$

A Cautious Note

Suppose that the posterior is $\pi\left(\theta, \tau \mid x\right)$, where θ is the parameter of interest. The mode of $\pi\left(\theta, \tau \mid x\right)$ may not equal the marginal posterior mode, the mode of $\pi\left(\theta \mid x\right)$.

Example

Consider the normal-inverse-gamma model, where the posterior is

$$\mu \mid \sigma^2, x \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \qquad \sigma^2 \mid x \sim \text{InvGamma}\left(a_n,\ b_n\right),$$

where μ_n, a_n, and b_n are known constants. Find the joint and marginal MAPs.

One More Issue: Existence

Example

Suppose that we observe iid $X_i \mid \theta \sim N(\theta, 1)$. The prior of θ is a mixture normal

$$\pi(\theta) = pN\left(\mu, \sigma^2\right) + (1-p)\, N\left(-\mu, \sigma^2\right).$$

Find the mode of $\pi(\theta \mid x)$.

Regularized Estimator

The MAP estimator essentially maximizes $f(x \mid \theta)\pi(\theta)$ or

$$\log f(x \mid \theta) + \log \pi(\theta),$$

provided that the logarithms are well defined. Intuitively speaking, we maximize the log-likelihood $\log f(x \mid \theta)$, but the penalty term $\log \pi(\theta)$ cannot be too big.

Suppose that data are generated from $X_i \mid \theta = \theta_0 \sim f(x \mid \theta_0)$, $i = 1, ..., n$.

- If $n^{-1} \log \pi(\theta) \to 0$ as $n \to \infty$, we should expect the MAP and the MLE to share similar properties.

One Difference Between MLE and MAP

Theorem (Invariance of MLE)

Let $\hat{\theta}_{ML}$ be the MLE of θ. Then, $g\left(\hat{\theta}_{ML}\right)$ is the MLE of $g\left(\theta\right)$ for any $g\left(\cdot\right)$.

However, MAP is not invariant with respect to reparametrization.

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. Let the prior be $\theta \sim \text{Beta}\left(a_0, b_0\right)$, where $a_0 > 1$ and $b_0 > 1$.

1. Find the MAP estimator of θ.
2. Find the MAP estimator of $\eta = \theta / \left(1 - \theta\right)$.

Posterior Mean

An alternative to MAP is the posterior mean

$$\hat{\theta}_{\text{Mean}} = \mathrm{E}\left[\theta \mid x\right].$$

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. Let the prior be $\theta \sim \text{Beta}\left(a_0, b_0\right)$, where $a_0 > 1$ and $b_0 > 1$.

- The posterior is $\text{Beta}\left(a_0 + x, b_0 + n - x\right)$.
- Hence, $\hat{\theta}_{\text{Mean}} = \frac{a_0+x}{a_0+b_0+n}$.

Posterior Mean and L_2 Loss

Consider the weighted L_2 loss

$$L_W(\theta, d) = (\theta - d)^T W (\theta - d),$$

where W is a $p \times p$ positive definite matrix and d is an estimator of θ using x.

Theorem

Suppose that there exists an estimator d such that

$$E[L_W(\theta, d) \mid x] = \int L_W(\theta, d)\, \pi(\theta \mid x)\, d\theta < \infty,$$

Then, the posterior mean minimizes $E[L_W(\theta, d) \mid x]$, where W does not depend on θ.

Posterior Mean versus MAP

Suppose that the posterior is $\pi\left(\theta,\tau \mid x\right)$, where θ is the parameter of interest.

- The mode of $\pi\left(\theta,\tau \mid x\right)$ may not equal the marginal posterior mode, the mode of $\pi\left(\theta \mid x\right)$.
- But the marginal posterior mean is the same as the joint posterior mean.

Example

Consider the normal-inverse-gamma model, where the posterior is

$$\mu \mid \sigma^2, x \sim N\left(\mu_n, \quad \frac{\sigma^2}{\lambda_0 + n}\right) \qquad \sigma^2 \mid x \sim \text{InvGamma}\left(a_n,\ b_n\right),$$

where μ_n, a_n, and b_n are known constants. The posterior mean is the mean of $\text{InvGamma}\left(a_n,\ b_n\right)$.

Posterior Mean versus MAP

To obtain the closed form expression of $\mathrm{E}[\theta \mid x]$, we need the normalizing constant of $\pi(\theta \mid x)$.

- The MAP estimator only requires the kernel $f(x \mid \theta)\pi(\theta)$. We can skip the integration step to obtain $m(x)$.
- Even we know $m(x)$, the integral to get $\mathrm{E}[\theta \mid x]$ may not be tractable.

But if we can sample from $\pi(\theta \mid x)$, we don't need to compute $m(x)$ nor evaluate the integral for $\mathrm{E}[\theta \mid x]$.

- Suppose that we have a sample θ_1, ..., θ_m from $\pi(\theta \mid x)$, then we can approximate the posterior mean by

$$\frac{1}{m}\sum_{j=1}^{m}\theta_j.$$

Posterior Mean versus MAP

It can even happen that the posterior mean does not exist, even though the posterior is proper.

Example

Let X_1, ..., X_n be iid from a two parameter Weibull distribution

$$f\left(x \mid \theta, \beta\right) = \frac{\beta x^{\beta-1}}{\theta^{\beta}} \exp\left\{-\left(\frac{x}{\theta}\right)^{\beta}\right\}, \quad x > 0,\ \theta > 0,\ \beta > 0.$$

Consider the proper priors

$$\begin{aligned} \pi\left(\theta \mid \beta\right) &= \frac{\beta b_0^{a_0}}{\Gamma\left(a_0\right)} \frac{1}{\theta^{a_0\beta+1}} \exp\left(-\frac{b_0}{\theta^{\beta}}\right), \text{ "InvGamma" prior} \\ \pi\left(\beta\right) &= \frac{d_0^{c_0}}{\Gamma\left(c_0\right)} \beta^{c_0-1} \exp\left(-d_0\beta\right). \text{ Gamma prior} \end{aligned}$$

With probability 1, the posterior mean of θ^k does not exist for any $k > 0$.

Posterior Mean versus MAP

It can happen that the likelihood involves intractable integrals. Hence, the MAP is not easy to obtain but we can sample easily from the posterior.

Example

Suppose that $Y_{ij} \mid Z_i, \beta, \lambda \sim \text{Bernoulli}\left(p_{ij}\right)$, $i = 1, ..., n$, $j = 1, ..., k$, where

$$p_{ij} = \frac{\exp\left(\beta_j + \lambda z_i\right)}{1 + \exp\left(\beta_j + \lambda z_i\right)}.$$

But we only observe $\{Y_{ij}\}$.

Invariance of Posterior Mean

The posterior mean is not invariant with respect to reparametrization either.

Example

Suppose that we observe one observation from $X \mid \theta \sim \text{Binomial}\,(n, \theta)$. Let the prior be $\theta \sim \text{Beta}\,(a_0, b_0)$, where $a_0 > 1$ and $b_0 > 1$.

1. Find the posterior mean of θ.
2. Find the posterior mean of $\eta = \theta / (1 - \theta)$.

Predict New Value

In frequentist statistics, the prediction of a new observation z after observing x is

$$\hat{z}\left(x\right) = \int zf\left(z \mid x, \hat{\theta}\right) dz.$$

In Bayesian statistics, the predictive distribution of a new observation z after observing x is

$$f\left(z \mid x\right) = \int f\left(z \mid x, \theta\right) \pi\left(\theta \mid x\right) d\theta.$$

A predictor can be the predictive mean

$$\hat{z}\left(x\right) = \int zf\left(z \mid x\right) dz,$$

or the predictive mode $\max\limits_{z} f\left(z \mid x\right)$.

Derive Predictive Distribution

Example

Consider an iid sample $(X_1, ..., X_n)$ from Poisson (θ). The prior of θ is Gamma (a_0, b_0) with density

$$\pi(\theta) = \frac{b_0^{a_0}}{\Gamma(a_0)} \theta^{a_0 - 1} \exp(-b_0 \theta).$$

1. Find the posterior $\pi(\theta \mid x)$.
2. Let z be a future value. Find the predictive distribution $f(z \mid x)$.
3. Propose a predictor of z.

Multiple Linear Regression

A multiple linear regression is

$$Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, ..., n,$$

where Y_i is the response, x_i is the vector of covariates (or regressors, or features), and β is the vector of unknown regression parameter.

In matrix notation, the model is

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}.$$

Some examples are:

1. Y is apartment price, $\boldsymbol{Z}$ includes crime rate, number of rooms, size of the apartment, year of construction, etc.
2. Y is waste water flow rate, $\boldsymbol{Z}$ includes temperature, precipitation, date of the year, time, etc.

Normal Linear Model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

The usual assumptions are

1. $\mathrm{E}\left[\epsilon \mid X\right] = 0$,
2. $\mathrm{Var}\left(\epsilon \mid X\right) = \Sigma$, where $\Sigma > 0$.

A typical assumption is $\Sigma = \sigma^2 I_n$, where I_n is an $n \times n$ identity matrix.

The ordinary least squares (OLS) estimator of β minimizes $\left(y - X\beta\right)^T \left(y - X\beta\right)$, and the minimizer is

$$\hat{\beta}_{\mathrm{OLS}} = \left(X^T X\right)^{-1} X^T y.$$

Normal Linear Model

In the normal linear model, we further assume that ϵ is normal: $\epsilon \mid X \sim N_n(0, \Sigma)$. Hence,

$$Y \mid X, \beta, \Sigma \quad \sim \quad N(X\beta,\ \Sigma).$$

The likelihood function is

$$f(y \mid X, \beta, \Sigma) \quad = \quad \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(y - X\beta)^T \Sigma^{-1} (y - X\beta)\right\}$$

If $\Sigma = \sigma^2 I$, the the MLE of β coincides with the OLS estimator:

$$\hat{\beta}_{\text{ML}} \quad = \quad \left(X^T X\right)^{-1} X^T y.$$

For notation simplicity, we will treat X as fixed and drop it from conditioning.

Bayesian Linear Model: Known Σ

Suppose that Σ is completely known, i.e., β is the only unknown parameter.

Result

The conjugate prior for β is $N_p\left(\mu_0,\, \Lambda_0^{-1}\right)$. The posterior is $\beta \mid y \sim N\left(\mu_n,\, \Lambda_n^{-1}\right)$, where

$$\begin{aligned} \Lambda_n &= \Lambda_0 + X^T\Sigma^{-1}X, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0\mu_0 + X^T\Sigma^{-1}y\right). \end{aligned}$$

Suppose that we observe a new x_0 and want to predict the new y_0. If $y_0 \mid \beta \sim N\left(x_0^T\beta,\, \sigma^2\right)$, where σ^2 is known, then the predictive distribution is

$$y_0 \mid y \sim N\left(x_0^T\mu_n,\, \sigma^2 + x_0^T\Lambda_n^{-1}x_0\right).$$

Ridge Regression

Suppose that $Y \mid \beta \sim N_n\left(X\beta, \sigma^2 I_n\right)$, where σ^2 is known. Let $\mu_0 = 0$ and $\Lambda_0 = \frac{\lambda}{\sigma^2} I_n$, that is

$$\beta \;\sim\; N_p\left(0,\, \frac{\sigma^2}{\lambda} I_n\right).$$

The posterior is

$$\beta \mid y \;\sim\; N_p\left(\left(X^T X + \lambda I_n\right)^{-1} X^T y,\, \frac{X^T X + \sigma^2 I_n}{\sigma^2}\right).$$

The posterior mean and MAP give the ridge regression estimator that minimizes

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \left(y - X\beta\right)^T \left(y - X\beta\right) + \lambda \beta^T \beta \\ &= \arg\max_{\beta} -\frac{1}{2\sigma^2}\left(y - X\beta\right)^T \left(y - X\beta\right) - \frac{1}{2\sigma^2/\lambda}\beta^T\beta.\end{aligned}$$

Laplace Prior

Consider the independent Laplace prior

$$\beta_i \quad \overset{iid}{\sim} \quad \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right).$$

The posterior satisfies

$$\pi\left(\beta \mid y\right) \quad \propto \quad \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right]\right\}.$$

The MAP gives the lasso regression estimator that minimizes

$$\begin{aligned}\hat{\beta} \quad &= \quad \arg\min_{\beta}\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \\ &= \quad \arg\max_{\beta} -\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right].\end{aligned}$$

Bayesian Linear Model: Unknown σ^2

Suppose that $\Sigma = \sigma^2 I_n$, but σ^2 is unknown. The parameter is $\theta = \left(\beta, \sigma^2\right)$.

- The likelihood is

$$\begin{aligned} f\left(y \mid \beta, \sigma^2\right) &= \frac{\exp\left\{-\frac{1}{2}\left(y - X\beta\right)^T\left(\sigma^2 I_n\right)^{-1}\left(y - X\beta\right)\right\}}{\left(2\pi\right)^{n/2}\sqrt{\det\left(\sigma^2 I_n\right)}} \\ &\propto \frac{1}{\left(\sigma^2\right)^{n/2}}\exp\left\{-\frac{\beta^T X^T X\beta - 2y^T X\beta}{2\sigma^2}\right\}. \end{aligned}$$

- The conjugate prior is

$$\begin{aligned} \beta \mid \sigma^2 &\sim N_p\left(\mu_0, \sigma^2\Lambda_0^{-1}\right), \\ \sigma^2 &\sim \text{InvGamma}\left(a_0, b_0\right), \end{aligned}$$

  a normal-inverse-gamma distribution.

Normal-Inverse-Gamma Distribution

Definition

A random vector $X \in \mathbb{R}^p$ and a positive random scalar $\lambda > 0$ follow a normal-inverse-gamma (NIG) distribution if

$$X \mid \lambda \sim N_p\left(\mu,\ \lambda\Sigma\right),\ \text{and } \lambda \sim \operatorname{InvGamma}\left(a, b\right).$$

It is denoted by $(X, \lambda) \sim \operatorname{NIG}\left(a, b, \mu, \Sigma\right)$. The joint density is

$$f\left(x, \lambda\right) \quad = \quad c \exp\left\{-\frac{\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)}{2\lambda} - \frac{b}{\lambda}\right\} \frac{1}{\lambda^{a+p/2+1}},$$

where the constant c is given by

$$c \quad = \quad \frac{b^a}{\left(2\pi\right)^{p/2} \Gamma\left(a\right) \sqrt{\det\left(\Sigma\right)}}.$$

Marginal Distribution

A random vector $X \in \mathbb{R}^p$ follows a multivariate t-distribution $t_v\left(\mu, \Sigma\right)$, if its density is

$$f\left(x\right) = \frac{\Gamma\left(\frac{v+p}{2}\right)}{\Gamma\left(\frac{v}{2}\right) v^{p/2} \pi^{p/2} \sqrt{\det\left(\Sigma\right)}} \left[1 + \frac{1}{v}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right]^{-(v+p)/2},$$

where v is the degrees of freedom, $\mu = \mathrm{E}\left[X\right]$ for $v > 1$, and $\mathrm{var}\left(X\right) = \frac{v}{v-2}\Sigma$ for $v > 2$.

Result

For the NIG distribution $\left(X, \lambda\right) \sim \mathrm{NIG}\left(a, b, \mu, \Sigma\right)$, the marginal distributions are

$$\begin{aligned} \lambda &\sim \mathrm{InvGamma}\left(a,\ b\right), \\ X &\sim t_{2a}\left(\mu,\ \frac{b}{a}\Sigma\right). \end{aligned}$$

Posterior Distribution

Result

Under the conjugate prior, the posterior distribution is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \sigma^2\Lambda_n^{-1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n,\ b_n\right), \end{aligned}$$

where

$$\begin{aligned} \Lambda_n &= X^T X + \Lambda_0, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0\mu_0 + X^T y\right), \\ a_n &= \frac{n}{2} + a_0, \\ b_n &= b_0 + \frac{1}{2}\left(y^T y + \mu_0^T\Lambda_0\mu_0 - \mu_n^T\Lambda_n\mu_n\right). \end{aligned}$$

That is, $\left(\beta, \sigma^2\right) \mid y \sim \text{NIG}\left(a_n, b_n, \mu_n, \Lambda_n^{-1}\right)$.

Marginal Posterior of Normal Linear Model

Under the conjugate prior, the posterior distribution is

$$\begin{aligned}\beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right),\\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n,\ b_n\right),\end{aligned}$$

that is $\left(\beta, \sigma^2\right) \mid y \sim \text{NIG}\left(a_n, b_n, \mu_n, \Lambda_n^{-1}\right)$. Then,

$$\begin{aligned}\beta \mid y &\sim t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n}\Lambda_n^{-1}\right),\\ \sigma^2 \mid y &\sim \text{InvGamma}\left(a_n, b_n\right).\end{aligned}$$

Predictive Distribution

Result

Suppose that we observe a new x_0 and want to predict the new y_0. Assume that $y_0 \perp y \mid \beta, \sigma^2$. Under the conjugate prior, the predictive distribution is

$$y_0 \mid y \quad \sim \quad t_{2a_n}\left(x_0^T\mu_n, \, \frac{b_n}{a_n}\left(1 + x_0^T\Lambda_n^{-1}x_0\right)\right),$$

same expectation as σ^2 were known.

Ridge Regression Again

Suppose that $Y \mid \beta \sim N_n\left(X\beta, \sigma^2 I_n\right)$, where σ^2 is unknown. Let $\mu_0 = 0$ and $\Lambda_0 = \lambda I_n$, that is

$$\beta \mid \sigma^2 \sim N_p\left(\mu_0, \frac{\sigma^2}{\lambda} I_p\right), \qquad \sigma^2 \sim \text{InvGamma}\left(a_0, b_0\right).$$

The posterior satisfies

$$\beta \mid y, \sigma^2 \sim N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right), \qquad \beta \mid y \sim t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n} \Lambda_n^{-1}\right),$$

where

$$\mu_n = \left(X^T X + \lambda I_n\right)^{-1} X y$$

coincides with the ridge regression estimator.

Laplace Prior Again

Consider the independent Laplace prior

$$\beta_j \mid \sigma \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right).$$

The posterior satisfies

$$\pi\left(\beta, \sigma^2 \mid y\right) \propto \frac{\pi\left(\sigma^2\right)}{\left(\sigma^2\right)^{p+n/2}} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(y - X\beta\right)^T\left(y - X\beta\right) + \lambda \sum_{j=1}^{p} |\beta_j|\right]\right\}.$$

The MAP gives the lasso regression estimator.

Tuning Parameter

The tuning parameter λ is often selected using cross validation in ridge/lasso regression.

In Bayesian linear model, we can also treat λ as an unknown variable and use a prior for λ. A hierarchical setup can be

$$
\begin{aligned}
y \mid \beta, \sigma^2 &\sim N\left(X\beta, \sigma^2 I_n\right) \\
\beta \mid \sigma^2, \lambda &\sim N_p\left(0, \frac{\sigma^2}{\lambda} I_p\right) \\
\sigma^2 &\sim \text{InvGamma}\left(a_0, b_0\right), \\
\lambda &\sim \text{InvGamma}\left(c_0, d_0\right).
\end{aligned}
$$

That is, the prior is $\pi\left(\beta, \sigma^2, \lambda\right) = \pi\left(\beta \mid \sigma^2, \lambda\right) \pi\left(\sigma^2\right) \pi\left(\lambda\right)$.

Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) of the model

$$Y = X\beta + e, \quad e \mid X \sim N_n\left(0, \sigma^2 I_n\right)$$

is given by

$$\hat{\beta}_{\mathrm{ML}} = \left(X^T X\right)^{-1} X^T y.$$

Its sampling distribution is

$$\hat{\beta}_{\mathrm{ML}} \mid \sigma^2 \sim N_p\left(\beta,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

The Zellner's g-prior is given by $\beta \mid \sigma^2 \sim N_p\left(\mu_0,\ g\sigma^2 \left(X^T X\right)^{-1}\right)$, where the constant $g > 0$.

Posterior Distribution

Result

Under the g-prior $\beta \mid \sigma^2 \sim N_p\left(\mu_0,\ g\sigma^2\left(X^TX\right)^{-1}\right)$ and $\sigma^2 \sim \operatorname{InvGamma}(a_0, b_0)$, the posterior distribution is

$$\begin{aligned}\beta \mid y, \sigma^2 &\sim N\left(\mu_n,\ \frac{g}{g+1}\sigma^2\left(X^TX\right)^{-1}\right), \\ \sigma^2 \mid y &\sim \operatorname{InvGamma}\left(\frac{n}{2}+a_0,\ b_n\right),\end{aligned}$$

where

$$\begin{aligned}\mu_n &= \frac{1}{g+1}\mu_0 + \frac{g}{g+1}\left(X^TX\right)^{-1}X^Ty, \\ b_n &= b_0 + \frac{1}{2}\left(y^Ty - \frac{g}{g+1}y^TX\left(X^TX\right)^{-1}X^Ty\right) \\ &\quad + \frac{1}{2}\left(\frac{1}{g+1}\mu_0^TX^TX\mu_0 - \frac{2}{g+1}y^TX\mu_0\right).\end{aligned}$$

Detour: Gradient and Hessian of Linear Form

Consider the function

$$f\left(x\right) = a_1 x_1 + a_2 x_2,$$

where $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ is a column vector. The gradient is

$$\frac{\partial f\left(x\right)}{\partial x} = \begin{bmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

The Hessian matrix is

$$\frac{\partial^2 f\left(x\right)}{\partial x \partial x^T} = \begin{bmatrix} \partial^2 f / \partial x_1^2 & \partial^2 f / \partial x_1 \partial x_2 \\ \partial^2 f / \partial x_2 \partial x_1 & \partial^2 f / \partial x_2^2 \end{bmatrix} = 0_{2\times 2}.$$

Detour: Gradient and Hessian of Quadratic Form

Consider

$$\begin{aligned} f\left(x\right) &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2. \end{aligned}$$

The gradient is

$$\frac{\partial f\left(x\right)}{\partial x} = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + a_{21}x_2 \\ a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 \end{bmatrix}.$$

The Hessian matrix is

$$\frac{\partial^2 f\left(x\right)}{\partial x \partial x^T} = \begin{bmatrix} 2a_{11} & a_{12} + a_{21} \\ a_{12} + a_{21} & 2a_{22} \end{bmatrix}.$$

General Results for Linear and Quadratic Form

If $f\left(x\right)=a^{T}x$ with a and x being $p\times 1$ column vectors, then

$$\begin{aligned}\frac{\partial f\left(x\right)}{\partial x} &= a,\\ \frac{\partial^{2} f\left(x\right)}{\partial x\partial x^{T}} &= 0_{p\times p}.\end{aligned}$$

If $f\left(x\right)=x^{T}Ax$ with x being a $p\times 1$ column vector, then

$$\begin{aligned}\frac{\partial f\left(x\right)}{\partial x} &= \left(A+A^{T}\right)x,\\ \frac{\partial^{2} f\left(x\right)}{\partial x\partial x^{T}} &= A+A^{T}.\end{aligned}$$

Jacobian Matrix

Suppose that the output of $f(x)$ is a $m \times 1$ vector, where the input x is a $p \times 1$ vector. The Jacobian matrix of f is defined to be

$$\frac{\partial f(x)}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x^T} \\ \frac{\partial f_2(x)}{\partial x^T} \\ \vdots \\ \frac{\partial f_m(x)}{\partial x^T} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_p} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_p} \end{bmatrix}_{m \times p}.$$

Example: Compute Jacobian Matrix

Example

Find the Jacobian matrix of $f\left(x\right)=\begin{bmatrix} a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

- Note that

$$\frac{\partial f_1\left(x\right)}{\partial x}=\frac{\partial a_1x_1+a_2x_2}{\partial x}=\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \qquad \frac{\partial f_2\left(x\right)}{\partial x}=\frac{\partial b_1x_1-b_2x_2}{\partial x}=\begin{bmatrix} b_1 \\ -b_2 \end{bmatrix}.$$

- Hence, the Jacobian matrix is

$$\frac{\partial f\left(x\right)}{\partial x^T}=\begin{bmatrix} \frac{\partial f_1(x)}{\partial x^T} \\ \frac{\partial f_2(x)}{\partial x^T} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ b_1 & -b_2 \end{bmatrix}.$$

In general, we have

$$\frac{\partial Ax}{\partial x^T} = A.$$

Jeffreys Prior

Result

Consider the linear model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$. The Fisher information of the above model is

$$\mathcal{I}\left(\beta, \sigma^2\right) = \begin{bmatrix} \frac{1}{\sigma^2} X^T X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Hence, the Jeffreys prior is

$$\pi\left(\beta, \sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{p/2+1}},$$

and the independent Jeffreys prior is

$$\pi\left(\beta, \sigma^2\right) \propto \frac{1}{\sigma^2}.$$

Independent Jeffreys Prior

The independent Jeffreys prior for the linear model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$ is

$$\pi\left(\beta, \sigma^2\right) \quad \propto \quad \frac{1}{\sigma^2}.$$

Consider the change of variables $\beta = \beta$ and $\tau = \log \sigma^2$. Then,

$$\pi\left(\beta, \tau\right) \quad \propto \quad \frac{1}{\sigma^2}\left|\det\left(\frac{\partial\begin{bmatrix}\beta \\ \sigma^2\end{bmatrix}}{\partial\begin{bmatrix}\beta & \tau\end{bmatrix}}\right)\right| = 1.$$

Hence, the independent Jeffreys prior means that the prior is uniform on $\left(\beta, \log \sigma^2\right)$.

Posterior with Jeffreys Prior

Theorem

Consider the linear regression model $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$. *Let the prior be*

$$\pi\left(\beta, \sigma^2\right) \quad \propto \quad \left(\sigma^2\right)^{-m}.$$

The posterior is

$$\begin{aligned} \beta \mid \sigma^2, y &\sim N_p\left(\mu_n,\, \sigma^2\left(X^T X\right)^{-1}\right), \\ \sigma^2 \mid y &\sim InvGamma\left(\frac{n-p}{2} + m - 1,\, \frac{1}{2} y^T\left(I_n - H\right) y\right), \end{aligned}$$

where $\mu_n = \left(X^T X\right)^{-1} X^T y$ *and* $H = X\left(X^T X\right)^{-1} X^T$ *is the hat matrix.*

MLE versus Posterior

The previous theorem shows that

$$\beta - \mu_n \mid \sigma^2, y \quad \sim \quad N_p\left(0,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

If maximum likelihood is used to estimate, then the MLE is $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$ and

$$\hat{\beta} - \beta \mid \beta, \sigma^2 \quad \sim \quad N_p\left(0,\ \sigma^2 \left(X^T X\right)^{-1}\right).$$

Posterior Predictive Checks

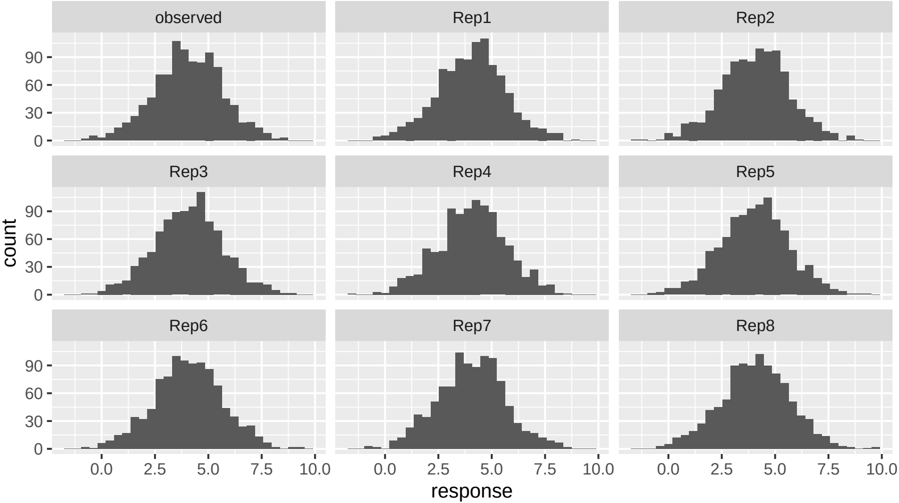
Posterior predictive check is a way to investigate whether our model can capture some relevant aspects of the data.

- We simulate data x_{sim} from the posterior predictive distribution

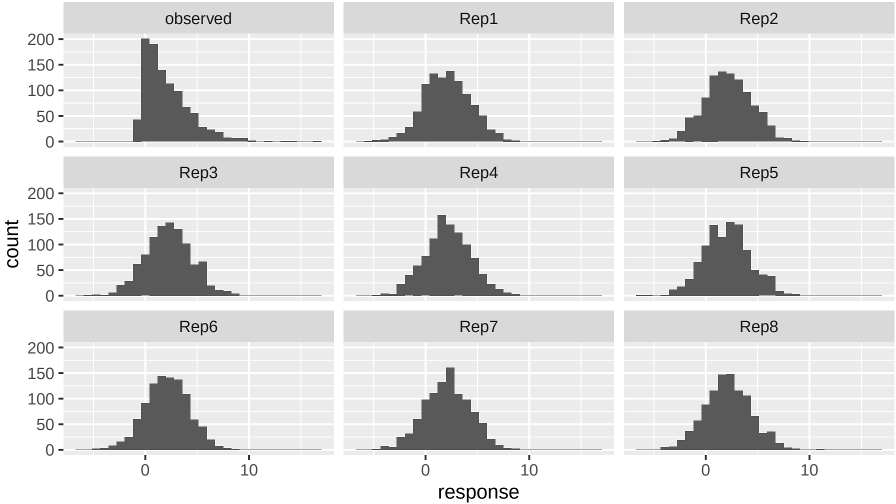
$$f\left(x_{\text{new}} \mid x\right) = \int f\left(x_{\text{new}} \mid x, \theta\right) \pi\left(\theta \mid x\right) d\theta.$$

- We can compare what our model predicts with the observed data, or compare statistics applied to the simulated data with the same statistics applied to the observed data.

Model 1


observed
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
count
90
60
30
0
0.0
2.5
5.0
7.5
10.0
response

Model 2


observed
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep7
Rep8
count
200
150
100
50
0
0
10
response

Gaussian Process

Definition

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Let f be a scalar-valued function. We denote a Gaussian process by

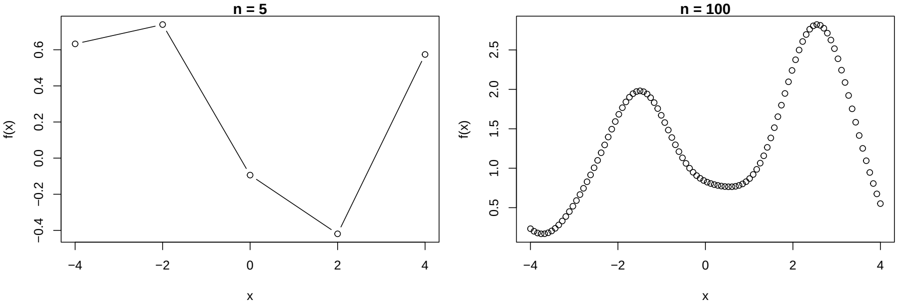
$$f\left(x\right) \quad \sim \quad \mathrm{GP}\left(m\left(x\right), k\left(x, x'\right)\right),$$

where $x \in \mathbb{R}^p$,

$$\begin{aligned} m\left(x\right) &= \mathrm{E}\left[f\left(x\right)\right], \text{ mean function} \\ k\left(x, x'\right) &= \operatorname{cov}\left(f\left(x\right), f\left(x'\right)\right). \text{ covariance function} \end{aligned}$$

Gaussian Process As Smooth Function

By the definition, the joint distribution of any finite $\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$ is multivariate normal. For a large enough n, the multivariate normal vector seems to produce a smooth function in x.



Gaussian Process Regression

Consider the Gaussian process regression model

$$y = f(x) + \epsilon,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$ and $f(x) \sim \mathrm{GP}\left(0, k\left(x, x'\right)\right)$. If we have observed n observations from this model, then

$$Y \mid \sigma^2 \sim N\left(0,\ K(X, X) + \sigma^2 I_n\right),$$

where

$$K(X, X) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{bmatrix}.$$

Recap: Conditional Distribution

Result: Conditional Distribution of Multivariate Gaussian Distribution

Suppose that $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_p \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$ such that $\Sigma_{22} > 0$. Then,

$$Y_1 \mid Y_2 = y_2 \quad \sim \quad N \left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} \left(y_2 - \mu_2 \right), \ \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right).$$

Predicted Value

Suppose that we want to predict the response value based on a new X_*. Then,

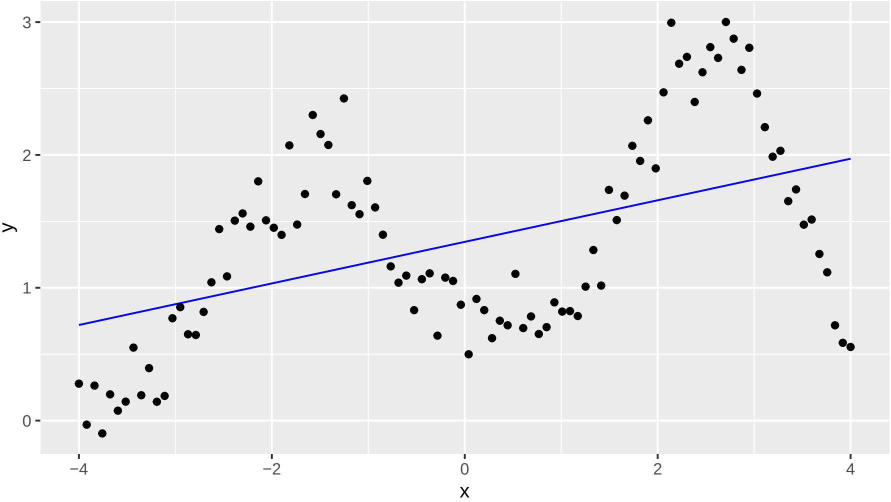
$$\begin{bmatrix} f\left(X_*\right) \\ Y \end{bmatrix} \mid \sigma^2 \quad \sim \quad N\left(0, \begin{bmatrix} K\left(X_*, X_*\right) & K\left(X_*, X\right) \\ K\left(X, X^*\right) & K\left(X, X\right) + \sigma^2 I_n \end{bmatrix}\right).$$

Hence, $f\left(X_*\right) \mid Y, \sigma^2$ is also Gaussian with

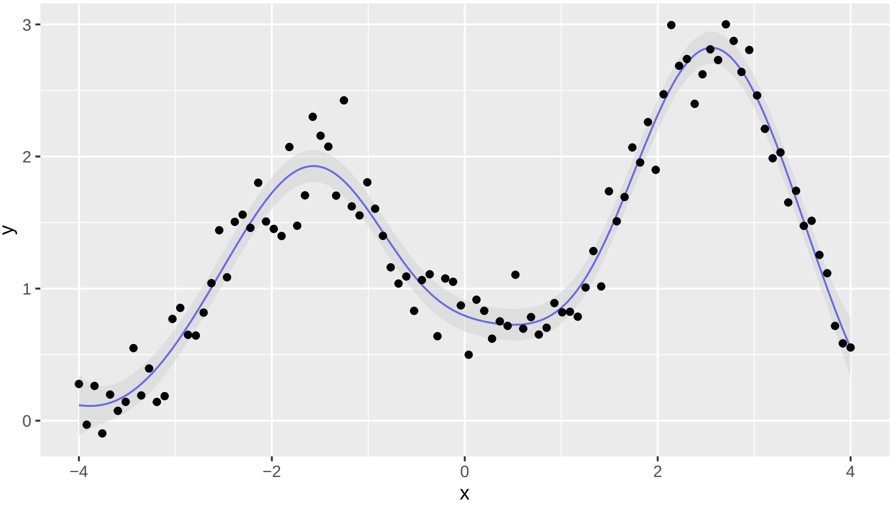
$$\begin{aligned} \text{mean} \qquad & K\left(X_*, X\right)\left[K\left(X, X\right) + \sigma^2 I_n\right]^{-1} y, \\ \text{covariance} \qquad & K\left(X_*, X_*\right) - K\left(X_*, X\right)\left[K\left(X, X\right) + \sigma^2 I_n\right]^{-1} K\left(X, X^*\right). \end{aligned}$$

The fitted function is then $K\left(X_*, X\right)\left[K\left(X, X\right) + \sigma^2 I_n\right]^{-1} y$.

Fitted Function Curve: $k(x, z) = x^T \Lambda_0^{-1} z$



Fitted Function Curve: $k(x,z) = \exp\left\{-\left\|x-z\right\|_2^2/2\right\}$



Bayesian Linear Model

Consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = x^T \beta,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- Under the conjugate prior $\beta \sim N_p\left(0,\ \Lambda_0^{-1}\right)$, the posterior is $\beta \mid y \sim N\left(\mu_n,\ \Lambda_n^{-1}\right)$, where

$$\mu_n \ = \ \left(\Lambda_0 + X^T X\right)^{-1} X^T y.$$

- Suppose that we observe a new x_0. and want to predict the new y_0. The predictive distribution is

$$y_0 \mid y \ \sim \ N\left(x_0^T \mu_n,\ \sigma^2 + x_0^T \Lambda_n^{-1} x_0\right).$$

Transform x

If we transform $x \in \mathbb{R}^p$ and obtain $\phi(x) \in \mathbb{R}^d$, then we can consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = \phi^T(x)\gamma,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- Under the conjugate prior $\gamma \sim N_d\left(0,\ \Omega_0^{-1}\right)$, the predictive distribution is

$$y_0 \mid y, \sigma^2 \quad \sim \quad N\left(\phi^T(x_0)\mu_n,\ \sigma^2 + \phi^T(x_0)\Omega_n^{-1}\phi(x_0)\right),$$

  where $\mu_n = \left(\sigma^2\Omega_0 + \phi^T(X)\phi(X)\right)^{-1}\phi^T(X)y$.
- The predictor is not linear in x_0 but linear in $\phi(x_0)$.

Kernel Function

A function $\kappa(x, z)$ is a kernel function if

1. it is symmetric, $\kappa(x, z) = \kappa(z, x)$,
2. the kernel matrix K with (i, j)th entry $\kappa(x_i, x_j)$ is positive semi-definite for all x_1, ..., x_n.

Example

Show that $\kappa(x, z) = x^T \Lambda_0^{-1} z$ is a kernel function for a symmetric Λ_0.

Rewrite Predictive Distribution

Let $\Phi = \phi\left(X\right) = \begin{bmatrix} \phi^T\left(x_1\right) \\ \vdots \\ \phi^T\left(x_n\right) \end{bmatrix}$. We can show that

$$\Omega_0^{-1}\Phi^T\left(\sigma^2 I_n + \Phi\Omega_0^{-1}\Phi^T\right)^{-1} = \left(\sigma^2\Omega_0 + \Phi^T\Phi\right)^{-1}\Phi^T.$$

Hence, the predictor from the predictive distribution is

$$\begin{aligned} \phi^T\left(x_0\right)\mu_n &= \phi^T\left(x_0\right)\left(\sigma^2\Omega_0 + \Phi^T\Phi\right)^{-1}\Phi^T y \\ &= \phi^T\left(x_0\right)\Omega_0^{-1}\Phi^T\left(\sigma^2 I_n + \Phi\Omega_0^{-1}\Phi^T\right)^{-1} y, \end{aligned}$$

where $\phi^T\left(x_0\right)\Omega_0^{-1}\Phi^T$ is a $1 \times n$ vector with elements $\left\{\phi^T\left(x_0\right)\Omega_0^{-1}\phi\left(x_i\right)\right\}$ and $\Phi\Omega_0^{-1}\Phi^T$ is a $n \times n$ matrix with elements $\left\{\phi\left(x_i\right)\Omega_0^{-1}\phi^T\left(x_j\right)\right\}$.

Example

Show that $\kappa\left(x, z\right) = \phi^T\left(x\right)\Omega_0^{-1}\phi\left(z\right)$ is a kernel function for a symmetric Ω_0.

Predictive Distribution Using Kernel Function

If $\kappa(x, z)$ is a kernel function, then we can find a function $\psi()$ such that $\kappa(x, z) = \psi^T(x)\psi(z)$.

- $\kappa(x, z) = \phi^T(x)\Omega_0^{-1}\phi(z) = \left[\Omega_0^{-1/2}\phi(x)\right]^T \Omega_0^{-1/2}\phi(z)$, where $\psi(x) = \Omega_0^{-1/2}\phi(x)$.

We can express the predictor from the predictive distribution as

$$\phi^T(x_0)\mu_n = K(x_0, X)\left[\sigma^2 I_n + K(X, X)\right]^{-1} y,$$

where

$$K(x_0, X) = \left\{\phi^T(x_0)\Omega_0^{-1}\phi(x_i)\right\} = \left\{\psi^T(x_0)\psi(x_i)\right\}$$

is a $1 \times n$ vector and

$$K(X, X) = \left\{\phi^T(x_i)\Omega_0^{-1}\phi(x_j)\right\} = \left\{\psi^T(x_i)\psi(x_j)\right\}$$

is a $n \times n$ matrix.

Kernel Trick

Our predictor $\phi^T(x_0)\,\mu_n$ depends on x only through the inner products $\psi^T(x)\,\psi(z)$ such as $\left\{\psi^T(x_0)\,\psi(x_i)\right\}$ and $\left\{\psi^T(x_i)\,\psi(x_j)\right\}$.

- Kernel trick is a commonly used trick to create new features from your original observed features, if our prediction depends on x only through inner products.
- By varying the kernel function, we obtain different sets of $\phi(x)$ and $\psi(x)$ as our new features.

Create New Feature

If $\kappa(x,z)$ is a kernel function, then we will have an eigen-decomposition

$$\kappa(x,z) = \sum_{m=1}^{\infty} \rho_m e_m(x) e_m(z),$$

for some eigenvalues ρ_k and eigenfunctions $e_m(x)$.

It can possibly be viewed as infinite new features have been created as

$$\kappa(x,z) = \sum_{m=1}^{\infty} \underbrace{\sqrt{\rho_m} e_m(x)}_{\text{new feature } \psi_m(x)} \underbrace{\sqrt{\rho_m} e_m(z)}_{\text{new feature } \psi_m(z)} .$$

Bayesian Regression and Gaussian Process

The predictive distribution $y_0 \mid y, \sigma^2$ is Gaussian with

$$\begin{aligned} \text{mean} \qquad & K\left(x_0, X\right)\left[\sigma^2 I_n + K\left(X, X\right)\right]^{-1} y, \\ \text{variance} \qquad & \sigma^2 + \phi^T\left(x_0\right)\left(\Omega_0 + \sigma^{-2}\Phi^T\Phi\right)^{-1}\phi\left(x_0\right). \end{aligned}$$

We can show that the variance is equivalent to

$$K\left(x_0, x_0\right) - K\left(x_0, X\right)\left(\sigma^2 I_n + K\left(X, X\right)\right)^{-1} K\left(X, x_0\right).$$

Recall that in Gaussian process regression, $f\left(x_0\right) \mid y, \sigma^2$ is also Gaussian with

$$\begin{aligned} \text{mean} \qquad & K\left(x_0, X\right)\left[\sigma^2 I_n + K\left(X, X\right)\right]^{-1} y, \\ \text{covariance} \qquad & K\left(x_0, x_0\right) - K\left(x_0, X\right)\left[K\left(X, X\right) + \sigma^2 I_n\right]^{-1} K\left(X, x_0\right). \end{aligned}$$

They are the same thing!

Prior on Function

Consider the linear regression model

$$y = f(x) + \epsilon, \qquad f(x) = \phi^T(x)\gamma,$$

where $\epsilon \mid \sigma^2 \sim N\left(0, \sigma^2\right)$.

- The conjugate prior $\gamma \sim N_d\left(0,\ \Omega_0^{-1}\right)$ implies that

$$f(x) \quad \sim \quad N\left(0,\ \phi^T(x)\,\Omega_0^{-1}\phi(x)\right).$$

  It can be viewed as the function has a Gaussian prior.

- The prior distribution of any set of function values satisfies

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} = \phi(X)\,\gamma \quad \sim \quad N_n\left(0,\ \phi(X)\,\Omega_0^{-1}\phi^T(X)\right).$$

Posterior on Function

The corresponding posterior is

$$\gamma \mid y, \sigma^2 \quad \sim \quad N\left(\mu_n,\, \Omega_n^{-1}\right),$$

where

$$\mu_n \quad = \quad \left(\sigma^2 \Omega_0 + \phi^T(X)\,\phi(X)\right)^{-1} \phi^T(X)\,y.$$

It can be viewed as the function has a Gaussian posterior

$$f(x) \mid y, \sigma^2 \quad \sim \quad N\left(\mu_n,\, \phi^T(x)\,\Omega_n^{-1}\phi(x)\right).$$

The predictive distribution is also Gaussian.

Bayesian Statistics
Bayesian Test

Shaobo Jin

Department of Mathematics

Hypothesis

Consider a statistical model $f(x \mid \theta)$ with $\theta \in \Theta$. We often want to investigate whether θ belongs to a subset of interest of Θ: $\theta \in \Theta_0$.

- For example, whether $\theta = 0$, or $\theta \leq \theta_0$.

The parameter space Θ is partitioned into two subsets,

- null hypothesis $H_0 : \theta \in \Theta_0$
- alternative hypothesis $H_1 : \theta \in \Theta_1$

such that $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

General Hypothesis

- The hypotheses are often formulated based on the restrictions on the parameter values, e.g., $\theta \in \Theta_0$.
- In general, a hypothesis formulates a class of statistical models.

Example

Suppose that we have randomly chosen n patients and want to analyze their blood samples in order to test drug resistance. Let X be the number of patients with positive test result. Two models are under consideration

$$\begin{aligned} H_0 : & \quad X \sim \text{Binomial}\left(n, p\right), \ p \sim \text{Beta}\left(a_0, b_0\right) \\ H_1 : & \quad X \sim \text{Poisson}\left(\lambda\right), \ \lambda \sim \text{Gamma}\left(a_1, b_1\right). \end{aligned}$$

Statistical Hypothesis Test

Definition

A nonrandomized test ϕ is a statistic from the sample space $\mathcal{X}$ to $\{0,1\}$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ (reject } H_0\text{)} \\ 0, & \text{if } x \in C_0, \text{ (do not reject } H_0\text{)} \end{cases}$$

where $\mathcal{X} = C_1 \cup C_0$ with $C_1 \cap C_0 = \emptyset$. A randomized test ϕ is a statistic from the sample space $\mathcal{X}$ to $[0,1]$:

$$\phi(x) = \begin{cases} 1, & \text{if } x \in C_1, \text{ (reject } H_0\text{)} \\ r, & \text{if } x \in C_=, \text{ (reject } H_0 \text{ with probability } r\text{)} \\ 0, & \text{if } x \in C_0, \text{ (do not reject } H_0\text{)} \end{cases}$$

where $\mathcal{X} = C_1 \cup C_= \cup C_0$ and $\{C_1, C_=, C_0\}$ are disjoint.

Type I Error and Type II Error

Statistical hypothesis testing is subject to errors.

	Truth	
Decision	H_0	H_1
H_0	Correct decision	Type II error
H_1	Type I error	Correct decision

- In general, a small Type I error probability yields a large Type II error probability, and vice versa.
- The theory of frequentist hypothesis testing is often built on the idea of most powerful test.
  - Among the tests that have low Type I error probability, we want to find the test that has the smallest Type II error.

Neyman-Pearson Test

Definition

Consider testing $H_0 : \mathrm{P}_0$ versus $H_1 : \mathrm{P}_1$. For $k \geq 0$, the randomized Neyman-Pearson test is

$$\phi\left(x\right) = \begin{cases} 1, & \text{if } f_0\left(x\right) < k f_1\left(x\right), \\ r, & \text{if } f_0\left(x\right) = k f_1\left(x\right), \\ 0, & \text{if } f_0\left(x\right) > k f_1\left(x\right), \end{cases}$$

where f_0 and f_1 are the density functions related to P_0 and P_1, respectively.

The Neyman-Pearson test is more powerful than any other test ϕ^* of level α, i.e., $\mathrm{E}\left[\phi^*\left(X\right) \mid H_0\right] \leq \alpha$.

Optimal Bayes Test

We can formulate the nonrandomized test as a decision problem. Suppose that the loss of the wrong decision is

	Truth	
Decision	H_0	H_1
H_0	0	a_1
H_1	a_0	0

such that $a_0 + a_1 > 0$.

Result

The optimal Bayes test that minimizes the posterior expected loss $\mathrm{E}\left[L \mid x\right]$ and the expected loss $\mathrm{E}\left[L\right]$ is

$$\phi\left(x\right) = \begin{cases} 1, & \text{if } \mathrm{P}\left(H_0 \mid x\right) < \frac{a_1}{a_0+a_1}, \\ 0, & \text{if } \mathrm{P}\left(H_0 \mid x\right) \geq \frac{a_1}{a_0+a_1}. \end{cases}$$

Bayes Test: Example

Example

1. Suppose that $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and $\theta \sim \text{Beta}\,(a, b)$. We are interested in testing

$$H_0 : \theta \geq \frac{1}{2}, \quad \text{versus} \quad H_1 : \theta < \frac{1}{2}.$$

2. Suppose that independent $X_i \mid \theta \sim N\left(\theta, \sigma^2\right)$ for $i = 1, ..., n$, where σ^2 is known. The prior is $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. We are interested in testing

$$H_0 : \theta \leq 0, \quad \text{versus} \quad H_1 : \theta > 0.$$

$0-1$ Loss

In the special case where $a_0 = a_1$, it is the same as the $0-1$ loss.

- Suppose that $\lambda = 0$ means that H_0 is the truth and $\lambda = 1$ that H_1 is the truth.
- The loss function is

$$L = \begin{cases} 1, & \text{if } \phi \neq \lambda, \\ 0, & \text{if } \phi = \lambda. \end{cases}$$

- The optimal Bayes test is equivalent to

$$\phi(x) = \begin{cases} 1, & \text{if } \mathrm{P}(H_0 \mid x) < \mathrm{P}(H_1 \mid x), \\ 0, & \text{if } \mathrm{P}(H_0 \mid x) \geq \mathrm{P}(H_1 \mid x). \end{cases}$$

Bayes Theorem for Simple Hypotheses

Consider two simple hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1.$$

The Bayes theorem yields

$$\mathrm{P}\left(H_k \mid x\right) = \frac{\mathrm{P}\left(H_k\right) f\left(x \mid \theta_k\right)}{\mathrm{P}\left(H_0\right) f\left(x \mid \theta_0\right) + \mathrm{P}\left(H_1\right) f\left(x \mid \theta_1\right)},$$

where $\mathrm{P}\left(H_k\right)$ is the prior probability that hypothesis H_k is true. Hence,

$$\frac{\mathrm{P}\left(H_0 \mid x\right) / \mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right) / \mathrm{P}\left(H_1\right)} = \frac{f\left(x \mid \theta_0\right)}{f\left(x \mid \theta_1\right)}.$$

Bayes Factor

In the case of simple hypotheses, comparing the odds

$$\frac{\mathrm{P}\left(H_0 \mid x\right)/\mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right)/\mathrm{P}\left(H_1\right)}$$

is equivalent to comparing the likelihood values. But we can still apply this ratio to more general cases.

Definition

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The Bayes factor is defined to be

$$B_{01}\left(x\right) = \frac{\mathrm{P}\left(H_0 \mid x\right)/\mathrm{P}\left(H_1 \mid x\right)}{\mathrm{P}\left(H_0\right)/\mathrm{P}\left(H_1\right)}.$$

Bayes Factor and Marginal Likelihood

Result

For $k \in \{0,1\}$, let $\pi_k(\theta)$ and $f_k(x \mid \theta)$ be the prior for θ and the likelihood under the hypothesis H_k, respectively. Let $\mathrm{P}(H_k)$ is the prior probability that hypothesis H_k is true. Then,

$$\underbrace{\frac{\mathrm{P}(H_0 \mid x)}{\mathrm{P}(H_1 \mid x)}}_{\text{posterior odds}} = \underbrace{\frac{\int_{\Theta_0} f_0(x \mid \theta)\,\pi_0(\theta)\,d\theta}{\int_{\Theta_1} f_1(x \mid \theta)\,\pi_1(\theta)\,d\theta}}_{\text{Bayes factor}} \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{prior odds}}.$$

Here,

$$m_k(x) = \int_{\Theta_k} f_k(x \mid \theta)\,\pi_k(\theta)\,d\theta$$

is the marginal likelihood under hypothesis H_k. Hence, the Bayes factor is also the ratio of marginal likelihoods.

Bayes Factor vs Likelihood Ratio Test

The Bayes factor is a ratio of marginal likelihoods

$$B_{01}(x) = \frac{\int\limits_{\Theta_0} f_0(x \mid \theta)\, \pi_0(\theta)\, d\theta}{\int\limits_{\Theta_1} f_1(x \mid \theta)\, \pi_1(\theta)\, d\theta}.$$

The likelihood ratio test computes the ratio of maximum likelihoods

$$\lambda(x) = \frac{\sup\limits_{\Theta_0} f(x \mid \theta)}{\sup\limits_{\Theta} f(x \mid \theta)} = \frac{f\left(x \mid \hat{\theta}_0\right)}{f\left(x \mid \hat{\theta}\right)},$$

where $\hat{\theta}_0$ is the MLE with the restriction $\theta \in \Theta_0$ and $\hat{\theta}$ is the MLE without such restriction.

Rule-of-Thumb

A large $B_{01}(x)$ indicates that the marginal likelihood under H_0 is higher than that under H_1. A rule-of-thumb to interpret the value of Bayes factor B_{10} (instead of B_{01}) is as follows.

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

A side note is that the marginal likelihood $m(x)$ is the omitted normalizing constant when we use $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$. We have to keep track all constants now!

Compute Bayes Factor: Simple H_0

Example

Compute Bayes factor.

1. Suppose that $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$ and $\theta \sim \text{Beta}\left(a, b\right)$. We are interested in testing

$$H_0 : \theta = \frac{1}{2}, \quad \text{versus} \quad H_1 : \theta \neq \frac{1}{2}.$$

   The prior under the alternative hypothesis is Uniform $[0, 1]$.

2. Suppose that $X_i, ..., X_n \mid \theta, \sigma^2$ be iid $N\left(\theta, \sigma^2\right)$, where both θ and σ^2 are unknown. We are interested in testing

$$H_0 : \theta = 0, \quad \text{versus} \quad H_1 : \theta \neq 0.$$

   The prior for $\theta \mid \sigma^2$ under the alternative hypothesis is $N\left(0, \sigma^2\right)$. The prior for σ^2 is σ^{-2}.

Compute Bayes Factor: Complicated Example

Example (Two-sample t test)

Suppose that we have two independent samples, $X_i \sim N\left(\mu_1, \sigma^2\right)$, $i = 1, ..., n_1$ and $Y_j \sim N\left(\mu_2, \sigma^2\right)$, $j = 1, ..., n_2$. We want to test whether their expectations are the same. We can reparametrize the distributions as

$$X_i \sim N\left(\mu + 2^{-1}\delta, \sigma^2\right) \qquad Y_j \sim N\left(\mu - 2^{-1}\delta, \sigma^2\right),$$

with parameters $\left(\mu, \delta, \sigma^2\right)$, where $\delta = \mu_1 - \mu_2$, $\mu = \left(\mu_1 + \mu_2\right)/2$. The hypotheses are

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0.$$

Let $\pi_0\left(\mu, \sigma^2\right) = \sigma^{-2}$ for both hypotheses and $\delta \mid \mu, \sigma^2 \sim N\left(0, \sigma_0^2\sigma^2\right)$ for H_1. Find the Bayes factor.

Bayes Factor and Optimal Bayes Test

Recall that the optimal Bayes test is

$$\phi(x) = \begin{cases} 1, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) < \frac{a_1}{a_0+a_1}, \\ 0, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) \geq \frac{a_1}{a_0+a_1}. \end{cases}$$

From the expression of the Bayes factor, we obtain

$$\mathrm{P}(\theta \in \Theta_0 \mid x) = \frac{B_{01}\mathrm{P}(\theta \in \Theta_0)}{\mathrm{P}(\theta \in \Theta_1) + B_{01}\mathrm{P}(\theta \in \Theta_0)}.$$

Hence, rejecting H_0 by the optimal Bayes test is equivalent to rejecting H_0 if

$$B_{01} < \frac{a_1\mathrm{P}(\theta \in \Theta_1)}{a_0\mathrm{P}(\theta \in \Theta_0)}.$$

Bayes Factor with Improper Prior

As long as the posterior is proper, we can use an improper prior in estimation. However, we need to be careful when using improper prior with Bayes factors.

- Suppose that we have one observation $X \sim N(\theta, 1)$ and consider the Jeffreys prior $\pi(\theta) \propto 1$.
- We want to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.
- The Bayes factor is

$$\begin{aligned} B_{10} &= \frac{\int_{\theta \in \Theta_1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \cdot 1 d\theta}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)} \\ &= \sqrt{2\pi} \exp\left(\frac{x^2}{2}\right) \geq \sqrt{2\pi} = 2.5066, \quad \forall x. \end{aligned}$$

- The Bayes factor is biased towards favoring H_1.

Jeffreys-Lindley Paradox

For the same example as above, consider the prior $\theta \sim N\left(0, \sigma_0^2\right)$.

- Intuitively speaking, as σ_0^2 increases, the prior becomes less informative.
- As $\sigma_0^2 \to \infty$, we should approximate an uninformative prior.
- The Bayes factor satisfies

$$B_{01} = \sqrt{\sigma_0^2 + 1} \exp\left(\frac{x^2}{2\left(\sigma_0^2 + 1\right)} - \frac{x^2}{2}\right) \to \infty,$$

  if $\sigma_0^2 \to \infty$, for any fixed x. Hence, it favors H_0 instead.

The Jeffreys-Lindley paradox says that the test based on an improper prior cannot be approximated by tests based on priors with increasing variances.

Training Sample

The intrinsic Bayes factor is a possible way out from the problems associated with improper priors.

Definition

Given an improper prior π, a sample $(x_1, ..., x_n)$ is called a training sample if the corresponding posterior $\pi(\theta \mid x_1, ..., x_n)$ is proper. The sample is a minimal training sample if no subsample is a training sample.

Example

Suppose that we have an iid sample $(x_1, ..., x_n)$ from $N(\mu, \sigma^2)$. Consider the Jeffreys prior.

1. If μ is unknown, but σ^2 is known, then the minimal training sample size is 1.
2. If both μ and σ^2 are unknown, then the minimal training sample size is 2.

Intrinsic Bayes Factor

The idea is to

1. use a training sample to produce a proper posterior from an improper prior,
2. use the resulting posterior as if it were a proper prior for the rest of the sample.

Definition

Let $x_{(l)}$ be a training sample and $x_{-(l)}$ be the rest of the sample. The intrinsic Bayes factor is

$$B_{01}^{I} = \frac{\int_{\Theta_0} f_0\left(x_{-(l)} \mid \theta\right) \pi_0\left(\theta \mid x_{(l)}\right) d\theta}{\int_{\Theta_1} f_1\left(x_{-(l)} \mid \theta\right) \pi_1\left(\theta \mid x_{(l)}\right) d\theta}.$$

Equivalent Representation of Intrinsic Bayes Factor

Result

Suppose that we have an independent sample $(x_1, ..., x_n)$. The intrinsic Bayes factor can be written as

$$B_{01}^{I}(x) = \underbrace{\frac{\int\limits_{\Theta_0} f_0(x \mid \theta)\,\pi_0(\theta)\,d\theta}{\int\limits_{\Theta_1} f_1(x \mid \theta)\,\pi_1(\theta)\,d\theta}}_{B_{01}(x)} \underbrace{\frac{\int_{\Theta_1} f_0(x_l \mid \theta)\,\pi_1(\theta)\,d\theta}{\int_{\Theta_0} f_1(x_l \mid \theta)\,\pi_0(\theta)\,d\theta}}_{B_{10}\left(x_{(l)}\right)}.$$

The choice of training sample can influence the Bayes factor.

- The training sample should be chosen as small as possible, e.g., minimal training sample.
- Since we split the data into two parts, we avoid using the same data twice.

Credible Set

Definition

A set $C(x)$ is a α-credible set if the posterior distribution satisfies

$$\mathrm{P}\left(\theta \in C(x) \mid x\right) \quad \geq \quad 1-\alpha, \quad \alpha \in [0,1].$$

1. It is highest posterior density (HPD) if it can be written as

   $$\left\{\theta : \ \pi\left(\theta \mid x\right) > k_\alpha\right\} \subseteq C(x) \subseteq \left\{\theta : \ \pi\left(\theta \mid x\right) \geq k_\alpha\right\},$$

   where k_α is the largest bound such that

   $$\mathrm{P}\left(\theta \in C(x) \mid x\right) \quad \geq \quad 1-\alpha.$$

2. It is an equal tailed credible interval if the lower and upper bounds satisfy

   $$\mathrm{P}\left(\theta \leq L(x) \mid x\right) \quad = \quad \mathrm{P}\left(\theta \geq U(x) \mid x\right) = \alpha/2.$$

Find Credible Set

If the posterior is a continuous distribution with density $\pi\left(\theta \mid x\right)$, then the HPD credible set is

$$C\left(x\right) = \left\{\theta :\ \pi\left(\theta \mid x\right) \geq k_{\alpha}\right\}$$

such that $\mathrm{P}\left(\theta \in C\left(x\right) \mid x\right) \geq 1-\alpha$.

Example

Find the credible set.

1. Let X_1, ..., X_n be iid $N\left(0, \sigma^2\right)$. We assume that the prior of σ^2 is $\mathrm{InvGamma}\left(a_0, b_0\right)$.
2. Let X_1, ..., X_n be iid $N\left(\theta, 1\right)$. We assume that the prior of θ is a normal mixture of $N\left(m_1, \tau_1^2\right)$ and $N\left(m_2, \tau_2^2\right)$.

Some Remarks

To consider the HPD credible sets is motivated by the fact that they minimize the volume among α-credible sets.

- The equal tailed credible interval is easy to work with but may contain regions with low posterior.
- The HPD credible set is not guaranteed to be an interval.

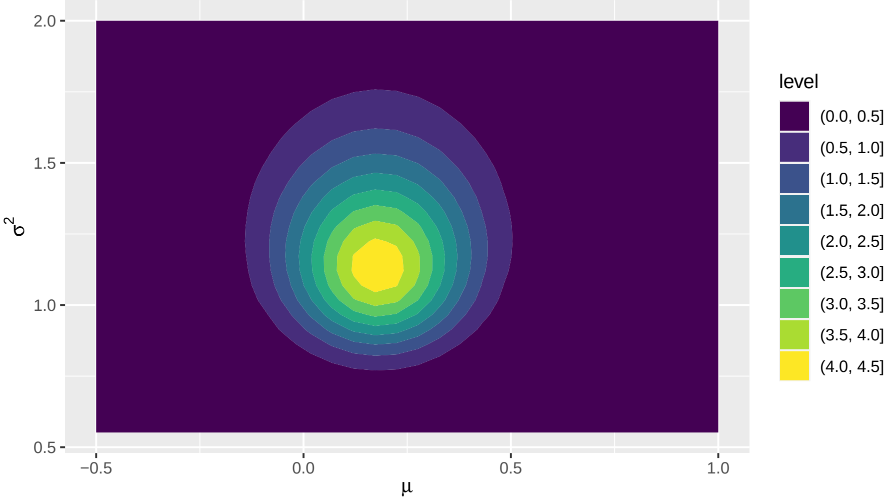
Improper priors can be used to find credible sets, as long as the posterior is proper.

Example

Let X_1, ..., X_n be iid $N\left(\mu, \sigma^2\right)$. We consider the prior $\pi\left(\mu, \sigma^2\right) = \sigma^{-2}$. Find the simultaneous credible set and the marginal credible interval.

Contour Plot of Example

Suppose that we observe $n = 50, \bar{x} = 0.18$, and $\sum_{i=1}^{n} x_i^2 = 60.53$.



Recall: Posterior in Linear model

Consider the linear model

$$Y = X\beta + \epsilon, \qquad \epsilon \mid \sigma^2 \sim N_n\left(0,\ \sigma^2 I_n\right).$$

Under the conjugate prior,

$$\beta \mid \sigma^2 \sim N_p\left(\mu_0, \sigma^2 \Lambda_0^{-1}\right), \qquad \sigma^2 \sim \mathrm{InvGamma}\left(a_0, b_0\right)$$

the posterior is

$$\beta \mid y, \sigma^2, N\left(\mu_n,\ \sigma^2 \Lambda_n^{-1}\right) \qquad \sigma^2 \mid y \sim \mathrm{InvGamma}\left(a_n,\ b_n\right).$$

The marginal posterior of β is

$$\beta \mid y \ \sim\ t_{2a_n}\left(\mu_n,\ \frac{b_n}{a_n}\Lambda_n^{-1}\right).$$

Example: Credible set in Linear model

The credible interval for σ^2 can be easily obtained from the marginal posterior

$$\sigma^2 \mid y \quad \sim \quad \mathrm{InvGamma}\left(a_n,\, b_n\right).$$

Lemma

If a $p \times 1$ random vector $X \sim t_v\left(\mu, \Sigma\right)$, then

$$\frac{1}{p}\left(X-\mu\right)^T \Sigma^{-1}\left(X-\mu\right) \quad \sim \quad F\left(p, v\right).$$

The lemma suggests that a credible set for β is

$$\left\{\beta :\ \frac{1}{p}\left(\beta-\mu_n\right)^T\left(\frac{b_n}{a_n}\Lambda_n^{-1}\right)^{-1}\left(\beta-\mu_n\right) \leq F\left(1-\alpha;\, p, v\right)\right\}.$$

Bayesian Statistics
Computational Techniques

Shaobo Jin

Department of Mathematics

Laplace Approximation to Integral

Suppose that we want to approximate the integral

$$\int h\left(\theta\right)\exp\left\{-\ell\left(\theta\right)\right\}d\theta,$$

where θ has the dimension $d \times 1$, p is a known constant, and $\ell\left(\theta\right)$ and $h\left(\theta\right)$ are smooth functions. If $\ell\left(\theta\right)$ is uniquely minimized at $\hat{\theta}$ such that

$$\frac{\partial\ell\left(\hat{\theta}\right)}{\partial\theta} = 0, \qquad \frac{\partial^2\ell\left(\hat{\theta}\right)}{\partial\theta\partial\theta^T} > 0.$$

Then, the Laplace approximation to the above integral is

$$\left(2\pi\right)^{d/2}\sqrt{\det\left(\left[\frac{\partial^2\ell\left(\hat{\theta}\right)}{\partial\theta\partial\theta^T}\right]^{-1}\right)}\exp\left\{-\ell\left(\hat{\theta}\right)\right\}h\left(\hat{\theta}\right).$$

Laplace Approximation: Example

Example

Suppose that posterior is

$$\begin{aligned}\beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^n y_i}{n+1},\ \frac{\sigma^2}{n+1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2},\ 2+\frac{1}{2}\left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n+1}\right]\right),\end{aligned}$$

where $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. Approximate $\mathrm{E}\left[\beta \mid y\right]$ by Laplace approximation.

Error Analysis

Suppose that we can express $\ell\left(\theta\right)$ as $pq\left(\theta\right)$ such that the integral is

$$\int h\left(\theta\right)\exp\left\{-pq\left(\theta\right)\right\}d\theta,$$

where $h\left(\theta\right)$ and $q\left(\theta\right)$ do not include p, and $h\left(\theta\right)$ and $q\left(\theta\right)$ are smooth functions.

Let $H \overset{\text{def}}{=} \frac{\partial^2 q\left(\hat{\theta}\right)}{\partial\theta\partial\theta^T} > 0$. The Laplace approximation satisfies

$$\left(2\pi\right)^{d/2}\sqrt{\det\left(H^{-1}\left(\hat{\theta}\right)/p\right)}\exp\left\{-pq\left(\hat{\theta}\right)\right\}\left[h\left(\hat{\theta}\right)+O\left(p^{-1}\right)\right].$$

Ratio of Integrals

In practice, we often want to approximate a ratio of integrals such that

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

The naive approach is to approximate both the numerator and denominator separately by Laplace approximation and take the ratio of approximations. This yields

$$\int h(\theta)\pi(\theta \mid x)\,d\theta \;\approx\; h\left(\hat{\theta}\right),$$

which is not recommended.

Moment Generation Function

Consider the moment generation function

$$\begin{aligned} \mathrm{E}\left[\exp\left\{th\left(\theta\right)\right\} \mid x\right] &= \int \exp\left\{th\left(\theta\right)\right\}\pi\left(\theta|x\right)d\theta \\ &= \frac{\int \exp\left\{th\left(\theta\right)\right\} f\left(x \mid \theta\right)\pi\left(\theta\right)d\theta}{\int f\left(x \mid \theta\right)\pi\left(\theta\right)d\theta}. \end{aligned}$$

We apply the Laplace approximation both the denominator and numerator, and take the ratio of approximations. Using the property of the moment generation function,

$$\mathrm{E}\left[h\left(\theta\right) \mid x\right] = \left.\frac{d\log \mathrm{E}\left[\exp\left\{th\left(\theta\right)\right\} \mid x\right]}{dt}\right|_{t=0}.$$

Fully Exponential Laplace Approximation

Let

$$\ell\left(\theta,t\right) = -th\left(\theta\right)-\log f\left(x\mid\theta\right)-\log\pi\left(\theta\right).$$

The fully exponential Laplace approximation is

$$\mathrm{E}\left[h\left(\theta\right)\mid x\right] \approx h\left(\hat{\theta}\right)-\frac{1}{2}\frac{\partial}{\partial t}\log\left|\frac{\partial^{2}\ell\left(\tilde{\theta}\left(t\right),t\right)}{\partial\theta\partial\theta^{T}}\right|\Bigg|_{t=0},$$

where $\tilde{\theta}\left(t\right)$ maximizes $\ell\left(\theta,t\right)$ for a given t, and $\hat{\theta}=\tilde{\theta}\left(0\right)$.

Under the same assumptions as for the Laplace approximation, the error rate is $O\left(p^{-2}\right)$.

Expectation Under Posterior

For given data x, we often need to compute the posterior expected value of a function $h(\theta, x)$,

$$\mu(x) = \int h(\theta, x)\, \pi(\theta \mid x)\, d\theta.$$

However, it is not always the case that we can find the closed form expression of $\mu(x)$. Approximations are more often needed.

Suppose that we want to approximate

$$\mathrm{E}[h(x)] = \int h(x)\, f(x)\, dx,$$

where $f(x)$ is the density of random variable/vector X. A natural approximation is to approximate it by the sample mean

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Approximate Expectation by Sample Mean

Under mild conditions, the sample mean

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

has nice properties.

1. Unbiasedness: $\mathrm{E}\left[\bar{h}\right] = \mathrm{E}\left[h(x)\right]$ for any n.
2. Consistency: $\bar{h} \to \mathrm{E}\left[h(x)\right]$ in probability, as $n \to \infty$.
3. Strong consistency: $\bar{h} \to \mathrm{E}\left[h(x)\right]$ almost surely, as $n \to \infty$.
4. Asymptotic normality: $\sqrt{n}\left(\bar{h} - \mathrm{E}\left[h(x)\right]\right) \to N\left(0,\ \mathrm{Var}\left[h(x)\right]\right)$ in distribution, as $n \to \infty$.

The classic methods (e.g., independent Monte Carlo and importance sampling) have these properties.

Sample From Posterior

For given data x, suppose that we want to compute the posterior expected value

$$\mu\left(x\right) = \int h\left(\theta, x\right) \pi\left(\theta \mid x\right) d\theta.$$

If $\pi\left(\theta \mid x\right)$ is a well-known distribution such that we can easily sample from it, then we draw R independent samples from $\pi\left(\theta \mid x\right)$ and the independent Monte Carlo approximation is

$$\hat{\mu}^{\text{IMC}} = \frac{1}{n} \sum_{i=1}^{n} h\left(\theta_i, x\right).$$

Independent Monte Carlo: Example

Example

Suppose that posterior is

$$\begin{aligned}
\beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^n y_i}{n+1}, \frac{\sigma^2}{n+1}\right), \\
\sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2}, 2+\frac{1}{2}\left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n+1}\right]\right),
\end{aligned}$$

where $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. We want to approximate $\mathrm{E}\left[\beta \mid y\right]$ by independent Monte Carlo. In this example, we know the true value

$$\mathrm{E}\left[\beta \mid y\right] = \frac{\sum_{i=1}^n y_i}{n+1}.$$

Importance Distribution

It is common that it is not straightforward to sample directly from $\pi(\theta \mid x)$. Suppose that it is easy for us to sample directly from another distribution with density $g(\theta \mid x)$ such that $g(\theta \mid x) > 0$ whenever $h(\theta, x)\, \pi(\theta \mid x) \neq 0$.

- We can rewrite $\mu(x)$ as

$$\mu(x) = \int h(\theta, x) \frac{\pi(\theta \mid x)}{g(\theta \mid x)} g(\theta \mid x)\, d\theta = \mathrm{E}\left[h(\theta, x) \frac{\pi(\theta \mid x)}{g(\theta \mid x)} \mid x \right],$$

  where the expectation is taken with respect to $\theta \mid x \sim g(\theta \mid x)$.
- We call $g(\theta \mid x)$ an importance distribution or instrumental distribution.

Importance Sampling Approximation

The importance sampling approximation is

$$\hat{\mu}^{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) \frac{\pi\left(\theta_i \mid x\right)}{g\left(\theta_i \mid x\right)}.$$

Example

Suppose that posterior is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^{n} y_i}{n+1}, \frac{\sigma^2}{n+1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2}, 2+\frac{1}{2}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n+1}\right]\right), \end{aligned}$$

where $n = 20$, $\sum_{i=1}^{n} y_i = 40.4$, and $\sum_{i=1}^{n} y_i^2 = 93.2$. Approximate $\mathrm{E}\left[\sigma^2 \mid y\right]$ by importance sampling.

Normalizing Constant

Since we often derive the posterior using $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ by ignoring the normalizing constant, we cannot always evaluate $\pi(\theta \mid x)$.

- It is easy to evaluate $f(x \mid \theta)\pi(\theta)$, but not $m(x)$.
- We can rewrite μ as

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

- We can apply the importance sampling trick to both integrals:

$$\int h(\theta, x) f(x \mid \theta)\pi(\theta)\,d\theta = \mathrm{E}\left[h(\theta, x) \underbrace{\frac{f(x \mid \theta)\pi(\theta)}{g(\theta \mid x)}}_{\text{importance weight } w(\theta,x)}\right]$$

where $g(\theta \mid x) > 0$ whenever $\pi(\theta \mid x) \neq 0$, stronger than IS.

Normalized Importance Sampling

The importance sampling approximations to the numerator and denominator are

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} w\left(\theta_i, x\right) h\left(\theta_i, x\right) &= \frac{1}{n}\sum_{i=1}^{n} \frac{f\left(x \mid \theta_i\right) \pi\left(\theta_i\right)}{g\left(\theta_i \mid x\right)} h\left(\theta_i, x\right), \\ \frac{1}{n}\sum_{i=1}^{n} w\left(\theta_i, x\right) &= \frac{1}{n}\sum_{i=1}^{n} \frac{f\left(x \mid \theta_i\right) \pi\left(\theta_i\right)}{g\left(\theta_i \mid x\right)}.\end{aligned}$$

The ratio is the normalized importance sampling estimator

$$\hat{\mu}^{\text{NIS}} = \frac{\sum_{i=1}^{n} w\left(\theta_i, x\right) h\left(\theta_i, x\right)}{\sum_{i=1}^{n} w\left(\theta_i, x\right)}.$$

Normalized Importance Sampling: Example

We can even ignore the constants in $f\left(y \mid \theta\right) \pi\left(\theta\right)$ in normalized importance sampling.

Example

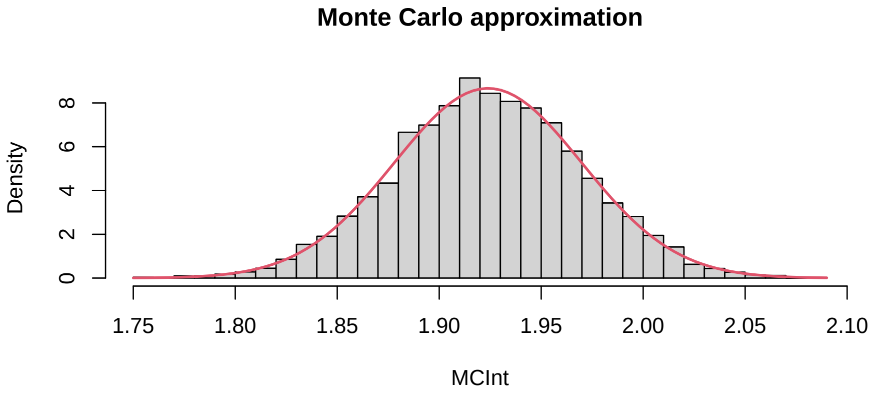
Consider an iid sample of size n from $Y \mid \beta, \sigma^2 \sim N\left(\beta, \sigma^2\right)$. The prior of σ^2 is $\text{InvGamma}\left(2, 2\right)$, and $\beta \mid \sigma^2$ is $N\left(0, \sigma^2\right)$. Then,

$$f\left(y \mid \theta\right) \pi\left(\theta\right) \quad \propto \quad \frac{\exp\left\{-\frac{(n+1)\beta^2 - 2\beta\sum_{i=1}^n y_i + 4 + \sum_{i=1}^n y_i^2}{2\sigma^2}\right\}}{\left(\sigma^2\right)^{(n+1)/2+3}}$$

We observe $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. Approximate $\text{E}\left[\sigma^2 \mid y\right]$ by normalized importance sampling.

Randomness

In independent Monte Carlo, importance sampling, and normalized importance sampling, we simulate random numbers from $\pi(\theta \mid x)$ or $g(\theta \mid x)$, then $\hat{\mu}$ is a random variable. This means that we can construct confidence interval for $\hat{\mu}$ using the central limit theorem.



Markov Chain Monte Carlo

We often want to get a sample from the posterior.

- If the posterior follows some well known distribution, we can generate a sample easily.
- If the posterior does not follow any well known distribution, the Markov Chain Monte Carlo (MCMC) is a very popular choice.

The idea of MCMC relies on the Markov property.

Definition

A Markov chain is a sequence of random variables X_i that satisfy the Markov property:

$$\mathrm{P}\left(X_{i+1} \in A \mid X_j = x_j, 0 \leq j \leq i\right) = \mathrm{P}\left(X_{i+1} \in A \mid X_i = x_i\right).$$

Transition Kernel

The transition kernel describes how the Markov chain moves from X_{n-1} to X_n.

- If $\{X_n\}$ is discrete, the transition kernel is a matrix K with elements $\mathrm{P}\left(X_n = y \mid X_{n-1} = x\right)$.
- If $\{X_n\}$ is continuous, the Markov property means that

$$\begin{aligned}
\mathrm{P}\left(X_n \in A \mid X_{n-1} = x, \cdots X_0\right) &= \int\limits_{y \in A} K\left(x, y\right) dy, \\
f\left(X_n = y \mid X_{n-1} = x, \cdots X_0\right) &= f\left(X_n = y \mid X_{n-1} = x\right) = K\left(x, y\right),
\end{aligned}$$

  where the transition kernel $K\left(x, y\right)$ is the conditional density of Y given $X = x$.

Stationary Distribution

Definition

The distribution p on Ω is a stationary distribution (or invariant distribution) of the Markov chain with the transition kernel K, if

$$\begin{aligned} \mathrm{P}\left(y\right) &= \sum_{x\in\mathcal{X}} \mathrm{P}\left(x\right) K\left(x,y\right), \quad \text{discrete case}, \\ f\left(y\right) &= \int_{x\in\mathcal{X}} f\left(x\right) K\left(x,y\right) dx, \quad \text{continuous case}, \end{aligned}$$

where P and f are not generic symbols.

- The stationary distribution means that if the initial state $X_0 \sim \pi\left(\theta \mid \text{data}\right)$, then $X_n \sim \pi\left(\theta \mid \text{data}\right)$ for all $n \geq 0$, the same distribution.

Long-Run Property

Theorem

Let $\pi()$ be the stationary distribution of the Markov chain. Under some regularity conditions,

$$\lim_{n\to\infty}\sup_{A}\left|P\left(X_n\in A\mid X_0=x\right)-\pi\left(A\right)\right| \quad = \quad 0, \text{ almost surely,}$$

regardless of the initial state $X_0=x$.

Since the limiting distribution does not depend on the initial state x, the marginal distribution of X_n is approximately the stationary distribution, after large enough iterations.

Choose the Transition Kernel

Our goal is to simulate data from $\pi(\theta \mid x)$. We need to choose the transition kernel K such that the stationary distribution is $\pi(\theta \mid x)$.

Fact

If $\pi(\theta \mid x)$ and $K(\theta, \theta^* \mid x)$ satisfies the detailed balance condition, i.e,

$$K(\theta, \theta^*)\,\pi(\theta \mid x) \quad = \quad K(\theta^*, \theta)\,\pi(\theta^* \mid x),$$

for any $\theta, \theta^* \in \Theta$, then $\pi(\theta \mid x)$ is the stationary distribution of the Markov chain with the transition kernel K.

Proposal Distribution

When we simulate random numbers from a Markov chain, we need a proposal distribution

$$T(\theta, \theta^*) \quad = \quad f(\theta^* \mid \theta).$$

Find a proposal distribution $T(\theta, \theta^*)$ that satisfies the detailed balance condition is difficult.

- So with probability $A(\theta, \theta^*)$ we let $\theta^{(n+1)} = \theta^*$ (accept), and probability $1 - A(\theta, \theta^*)$ we let $\theta^{(n+1)} = \theta$ (reject).
- For $\theta^{(n+1)} \neq \theta$, the transition is

  $$K(\theta, \theta^*) \quad = \quad T(\theta, \theta^*) A(\theta, \theta^*).$$

  Hence, we should seek A such that the detailed balance condition is fulfilled.

Deriving $A(\theta, \theta^*)$

The detailed balance condition is fulfilled, if we choose the acceptance probability to be

$$\begin{aligned} A(\theta, \theta^*) &= \lambda(\theta, \theta^*)\, \pi(\theta^* \mid x)\, T(\theta^*, \theta) \leq 1, \\ A(\theta^*, \theta) &= \lambda(\theta, \theta^*)\, \pi(\theta \mid x)\, T(\theta, \theta^*) \leq 1. \end{aligned}$$

The value λ that maximizes the probability $A(\cdot, \cdot) \leq 1$ is

$$\lambda(\theta, \theta^*) = \min\left\{ \frac{1}{\pi(\theta^* \mid x)\, T(\theta^*, \theta)}, \frac{1}{\pi(\theta \mid x)\, T(\theta, \theta^*)} \right\}.$$

Hence,

$$A(\theta, \theta^*) = \lambda(\theta, \theta^*)\, \pi(\theta^* \mid x)\, T(\theta^*, \theta) = \min\left\{ 1, \frac{\pi(\theta^* \mid x)\, T(\theta^*, \theta)}{\pi(\theta \mid x)\, T(\theta, \theta^*)} \right\}.$$

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm allows proposal distributions such that $T(\theta, \theta^*) > 0$ if and only if $T(\theta^*, \theta) > 0$.

Algorithm 1: Metropolis-Hastings Algorithm

```
Choose an initial state $\theta^{(0)}$ ;
for $t = 1$ in 1 : $n$ do
    Sample a candidate $\theta^*$ from $T\left(\theta^{(t)}, \theta \mid x\right)$ ;
    Calculate the ratio $R\left(\theta^{(t)}, \theta^*\right) = \frac{\pi(\theta^* | x) T\left(\theta^*, \theta^{(t)}\right)}{\pi\left(\theta^{(t)} | x\right) T\left(\theta^{(t)}, \theta^*\right)}$ ;
    Draw $U \sim \mathrm{U}[0,1]$ ;
    Update
        $\theta^{(t+1)} = \begin{cases} \theta^*, & \text{if } U \leq R\left(\theta^{(t)}, \theta^*\right), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$
end
```

Metropolis-Hastings Algorithm: Example

Since the ratio $R\left(\theta^{(t)}, \theta^{*}\right)$ includes $\frac{\pi(\theta^{*}|x)}{\pi\left(\theta^{(t)}|x\right)}$, we only need to know $\pi\left(\cdot \mid x\right)$ up to a normalizing constant.

Example

Consider an iid sample of size n from $Y \mid \beta, \sigma^2 \sim N\left(\beta, \sigma^2\right)$. The prior of σ^2 is $\text{InvGamma}\left(2, 2\right)$, and $\beta \mid \sigma^2$ is $N\left(0, \sigma^2\right)$. Then,

$$f\left(y \mid \theta\right) \pi\left(\theta\right) \quad \propto \quad \frac{\exp\left\{-\left[\left(n+1\right)\beta^2 - 2\beta\sum_{i=1}^{n} y_i + 4 + \sum_{i=1}^{n} y_i^2\right] / \left(2\sigma^2\right)\right\}}{\left(\sigma^2\right)^{(n+1)/2+3}}.$$

We observe $n = 20$, $\sum_{i=1}^{n} y_i = 40.4$, and $\sum_{i=1}^{n} y_i^2 = 93.2$. Obtain a sample from the posterior.

Detailed Balance: Symmetric Proposal

The Metropolis-Hastings algorithm allows asymmetric proposal distributions.

- If the proposal distribution is symmetric, i.e., $T(\theta, \theta^*) = T(\theta^*, \theta)$, then the Metropolis-Hastings algorithm reduces to the Metropolis algorithm.

Example

$\theta^* \mid \theta \sim N\left(\theta, \sigma^2\right)$ is symmetric, since

$$T(\theta, \theta^*) \quad = \quad \frac{1}{\sqrt{2\sigma^2}} \exp\left\{-\frac{(\theta - \theta^*)^2}{2\sigma^2}\right\}.$$

Metropolis Algorithm

Algorithm 2: Metropolis Algorithm

```
Choose an initial state $\theta^{(0)}$ ;
for $t = 1$ in 1 : $n$ do
    Sample a candidate $\theta^*$ from $T\left(\theta^{(t)}, \theta \mid x\right)$ ;
    Calculate the ratio $R\left(\theta^{(t)}, \theta^*\right) = \frac{\pi(\theta^* | x)}{\pi\left(\theta^{(t)} | x\right)}$ ;
    Draw $U \sim \mathrm{U}[0,1]$ ;
    Update
        $$\theta^{(t+1)} = \begin{cases} \theta^*, & \text{if } U \leq R\left(\theta^{(t)}, \theta^*\right), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$$
end
```

Some Examples of Metropolis-Hastings Algorithms

Many different MCMC algorithms differ mainly in how the candidate y is sampled.

- In the random-walk Metropolis algorithm, $\theta^* = \theta^{(t)} + \epsilon$, where ϵ is sampled from some distribution, e.g., Uniform $[-a, a]$, Normal, etc.
- In independence sampler, θ^* is sampled from $g(\cdot)$ that does not depend on $\theta^{(t)}$.
- The Langevin Metropolis-Hastings algorithm explores the shape of the posterior distribution by $\theta^* = \theta^{(t)} + d^{(t)} + \tau\epsilon$, where $\epsilon \sim N(0, I)$ and

$$d^{(t)} = \frac{\tau^2}{2} \frac{\partial \log \pi\left(\theta^{(t)} \mid x\right)}{\partial \theta}.$$

Gibbs Sampler: Conditioning

It can be the case that it is much easier to sample from the conditional distributions than using Metroplis-Hastings from the joint distribution of $\theta \in \Theta \subset \mathbb{R}^d$.

- Suppose that $\theta = (\theta_1, ..., \theta_p)$, where $\theta_i \in \mathbb{R}^{d_i}$.
- Let $\pi_{i|-i}(\theta_i \mid \theta_{-i}, x)$ be the conditional distribution of θ_i given θ_{-i} and x, where $\theta_{-i} = \begin{pmatrix}\theta_1 & \cdots & \theta_{i-1} & \theta_{i+1} & \cdots & \theta_p\end{pmatrix}$.

Algorithm 3: Basic Gibbs Sampler

```
Choose an initial state $\theta^{(0)}$ ;
for $t = 1$ in 1 : $n$ do
    for $i = 1$ in 1 : $p$ do
        Draw $\theta_i^{(t+1)} \sim \pi_{i|-i}\left(\theta_i \mid \theta_1^{(t+1)}, ..., \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, ..., \theta_p^{(t)}, x\right)$ ;
    end
end
```

Gibbs Sampler: Example

Example

Suppose that our data X_1, ..., X_n are iid from $N\left(\mu, \lambda^{-1}\right)$. The prior distributions of μ and λ are

$$\begin{aligned} \mu &\sim N\left(\mu_0, \lambda_0^{-1}\right), \\ \lambda &\sim \operatorname{Exp}\left(b_0\right). \end{aligned}$$

Use Gibbs sampler to sample random numbers from the posterior distribution of μ, λ.

Why Does Gibbs Sampler Work?

In order to show the Gibbs sampler generate random numbers from the desired stationary distribution, we only need to show

$$\pi\left(\theta^{*} \mid x\right) \quad = \quad \int K\left(\theta, \theta^{*}\right) \pi\left(\theta \mid x\right) d\theta,$$

where $\pi\left(\cdot \mid x\right)$ is not a generic symbol.

For simplicity, we consider $p=2$ and continuous posterior.

- The transition kernel $K\left(\theta, \theta^{*}\right)$ is

$$K\left(\left(\theta_1, \theta_2\right), \left(\theta_1^{*}, \theta_2^{*}\right)\right) \quad = \quad \pi_{1|2}\left(\theta_1^{*} \mid \theta_2, x\right) \pi_{2|1}\left(\theta_2^{*} \mid \theta_1^{*}, x\right).$$

- This transition kernel satisfies

$$\int\int K\left(\left(\theta_1, \theta_2\right), \left(\theta_1^{*}, \theta_2^{*}\right)\right) \pi\left(\theta_1, \theta_2 \mid x\right) d\theta_1 d\theta_2 \quad = \quad \pi\left(\theta_1^{*}, \theta_2^{*} \mid x\right).$$

Collapsed Gibbs Sampler

Suppose that θ can be partitioned into three groups of parameters $(\theta_1, \theta_2, \theta_3)$.

- The Gibbs sampler samples from the full conditional distributions $\theta_1^{(t+1)} \sim \pi\left(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, x\right)$, $\theta_2^{(t+1)} \sim \pi\left(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, x\right)$, and $\theta_3^{(t+1)} \sim \pi\left(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, x\right)$.

In collapsed Gibbs sampler, we can integrate out θ_3 analytically and work with $(\theta_1, \theta_2) \sim \pi(\theta_1, \theta_2 \mid x)$.

- We sample $\theta_1^{(t+1)} \sim \pi\left(\theta_1 \mid \theta_2^{(t)}, x\right)$ and $\theta_2^{(t+1)} \sim \pi\left(\theta_2 \mid \theta_1^{(t+1)}, x\right)$ by Gibbs sampler.
- We then sample $\theta_3^{(t+1)} \sim \pi\left(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, x\right)$.

Rao-Blackwell Theorem in Statistical Inference

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator of θ. Suppose that $T = T(X)$ is a sufficient statistic for θ. Then, $\theta^* = E\left[\hat{\theta} \mid T\right]$ is a uniformly minimum variance unbiased estimator of θ, i.e.,

$$Var\left(\hat{\theta}\right) \quad \geq \quad Var\left(\theta^*\right).$$

A weaker version is of the theorem is based on the low of total variance:

$$\operatorname{Var}(X) \quad = \quad \operatorname{Var}(\operatorname{E}[X \mid Y]) + \operatorname{E}(\operatorname{Var}[X \mid Y]) \geq \operatorname{Var}(\operatorname{E}[X \mid Y]).$$

Rao-Blackwellization

If we are interested in $\mathrm{E}\left[f\left(X,Y\right)\right]$, then

$$\mathrm{Var}\left(f\left(X,Y\right)\right) \quad \geq \quad \mathrm{Var}\left(\mathrm{E}\left[f\left(X,Y\right) \mid Y\right]\right).$$

That is, instead of simulating $\left(X_i, Y_i\right)$ to compute $n^{-1}\sum_{i=1}^{n} f\left(X_i, Y_i\right)$, we can simulate only Y_i and compute

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[f\left(X_i, Y_i\right) \mid Y_i\right].$$

This also suggests that we should compute as many analytical steps as possible before Monte Carlo approximation!

Rao-Blackwellization: Example

Example

Consider a Bayesian model, where $X_i \mid \mu, \lambda \sim N\left(\mu, \lambda^{-1}\right)$, $\mu \sim N\left(\mu_0, \lambda_0^{-1}\right)$, and $\lambda \sim \text{Gamma}\left(a_0, b_0\right)$. Then,

$$\begin{aligned}
\mu \mid \lambda, \text{data} &\sim N\left(\frac{\lambda_0\mu_0 + \lambda\sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}, \frac{1}{\lambda_0 + n\lambda}\right), \\
\lambda \mid \mu, \text{data} &\sim \text{Gamma}\left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i\mu + \frac{n}{2}\mu^2\right).
\end{aligned}$$

We want to approximate $\text{E}\left[\lambda \mid \text{data}\right]$.

Hamiltonian Monte Carlo

- The Metropolis algorithm and the Gibbs sampler often move too slowly through the target distribution when the dimension of the target distribution is high.
- Hamiltonian Monte Carlo (HMC) moves much quicker through the target distribution.
    - For each component in the target distribution, HMC adds a momentum variable and the proposal distribution largely depends on the momentum variable.
    - Both the component in the target distribution and the momentum are updated in the MCMC algorithm.

Hamiltonian Dynamics

The idea of HMC originates from the Hamiltonian dynamics in physics.

- The state of a system consists of the position $\theta \in \mathbb{R}^d$ and the momentum $\phi \in \mathbb{R}^d$ of same dimension.
- The Hamiltonian is a function of θ and ϕ, denoted by $H(\theta, \phi)$.
- The position and the momentum can change over time t. The change is described by the Hamilton's equations:

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i}, \quad \text{and} \quad \frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i},$$

for $i = 1, ..., d$.

Potential and Kinetic Energy

For HMC, the Hamiltonian is usually

$$H(\theta, \phi) = U(\theta) + V(\phi),$$

where $U(\theta) = -\log \pi(\theta \mid x)$ is called the potential energy and $V(\phi)$ is called the kinetic energy.

- We want to sample from $\pi(\theta \mid x)$. Hence, ϕ is artificial.
- We often let $\phi \sim N(0, M)$, independent of $\theta \mid x$, for a prespecified covariance matrix M, and $V(\phi)$ the negative log density of ϕ.
- The Hamilton's equations become

$$\frac{d\theta}{dt} = M^{-1}\phi, \quad \text{and} \quad \frac{d\phi}{dt} = \frac{\partial \log \pi(\theta \mid x)}{\partial \theta},$$

arranged as column vectors.

Augmentation

Since θ and ϕ are independent, their joint density is

$$\begin{aligned} f(\theta, \phi \mid x) = \pi(\theta \mid x) p(\phi \mid x) &= \exp\{-U(\theta) - V(\phi)\} \\ &= \exp\{-H(\theta, \phi)\}. \end{aligned}$$

We have augmented the problem from sampling θ from $\pi(\theta \mid x)$ to sampling (θ, ϕ) form $\exp\{-H(\theta, \phi)\}$.

1. We first sample ϕ from $N(0, M)$, independent of current θ.
   - Since $\phi \sim N(0, M)$, we already sample ϕ from the desired distribution.
2. We then sample θ, where the new state is proposed by Hamiltonian dynamics by solving the differential equations.

Solve Differential Equation

To solve the differential equations, we consider an approximation known as the leapfrog method. For some stepsize $\epsilon > 0$, we perform half-step updates as

$$\begin{aligned}
\phi\left(t+\frac{\epsilon}{2}\right) &= \phi\left(t\right)+\frac{\epsilon}{2}\frac{\partial\phi\left(t\right)}{\partial t}=\phi\left(t\right)+\frac{\epsilon}{2}\frac{\partial\log\pi\left(\theta\left(t\right)\mid x\right)}{\partial x},\\
\theta\left(t+\epsilon\right) &= \theta\left(t\right)+\epsilon\frac{\partial\theta\left(t+\frac{\epsilon}{2}\right)}{\partial t}=\theta\left(t\right)+\epsilon M^{-1}\phi\left(t+\frac{\epsilon}{2}\right),\\
\phi\left(t+\epsilon\right) &= \phi\left(t+\frac{\epsilon}{2}\right)+\frac{\epsilon}{2}\frac{\partial\phi\left(t+\epsilon\right)}{\partial t}=\phi\left(t+\frac{\epsilon}{2}\right)+\frac{\epsilon}{2}\frac{\partial\log\pi\left(\theta\left(t+\epsilon\right)\mid x\right)}{\partial\theta}.
\end{aligned}$$

Starting from $t=0$, we can get a trajectory at times $\{\epsilon, 2\epsilon, ..., L\epsilon\}$, and approximate the values for $\theta\left(L\epsilon\right)$ and $\phi\left(L\epsilon\right)$.

Leapfrog Method to Sample θ

Suppose that the current state is (θ, ϕ).

1. Update ϕ with a half-step update by
$$\phi \quad \leftarrow \quad \phi + \frac{\epsilon}{2} \frac{\partial \log \pi\left(\theta \mid x\right)}{\partial \theta}.$$
2. For $\ell = 1, ..., L-1$,
    1. Update the position: $\theta \leftarrow \theta + \epsilon M^{-1} \phi$.
    2. Update the momentum:
$$\phi \quad \leftarrow \quad \phi + \epsilon \frac{\partial \log \pi\left(\theta \mid x\right)}{\partial \theta}.$$
3. Make one last update on the position: $\theta \leftarrow \theta + \epsilon M^{-1} \phi$.
4. Make one last half-step update of the momentum
$$\phi \quad \leftarrow \quad \phi + \frac{\epsilon}{2} \frac{\partial \log \pi\left(\theta \mid x\right)}{\partial \theta}.$$

Metropolis Step

Suppose that the state after such L updates is (θ^*, ϕ^*). We negate the momentum and the new proposal state is $(\theta^*, -\phi^*)$.

- We determine whether to accept the proposal using the Metropolis algorithm, where the acceptance probability is

$$A\left(\left(\theta, \phi\right), \left(\theta^*, -\phi^*\right)\right) = \min\left\{1, \frac{\exp\left\{-H\left(\theta^*, -\phi^*\right)\right\}}{\exp\left\{-H\left(\theta, \phi\right)\right\}}\right\}.$$

- If the proposed state is accepted, then we accept θ^* as a new state for θ, but don't care about ϕ^*.
- No matter we accept or reject the proposal, we will draw a new momentum in the next iteration, independent of previous momentum.

Properties of HMC

Some crucial properties of the Hamiltonian dynamics for MCMC updates include

1. deterministic updates. The Hamiltonian dynamics is deterministic. After running the leapfrog loop L times, we always move the initial state (θ_0, ϕ_0) to the same proposal (θ^*, ϕ^*).
2. reversible. The mapping from the state at time t, denoted by $(\theta(t), \phi(t))$, to the state at time $t+s$, denoted by $(\theta(t+s), \phi(t+s))$, is one-to-one and has an inverse mapping. If we negate the momentum, we will come back from $(\theta(t+s), -\phi(t+s))$ to $(\theta(t), -\phi(t))$.
3. connection between momentum and position. The momentum is changed based on the position since

$$\frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i} = \frac{\partial \log \pi(\theta \mid x)}{\partial \theta_i}.$$

Tuning Parameters

Tuning of HMC can occur in several places such as

1. the distribution for the momentum,
2. the scaling factor ϵ,
3. the number of leapfrog steps L per iteration.

Some theory suggest that we can tune HMC such that the acceptance probability is around 65%.

No-U-Turn Sampler

The no-U-turn sampler (NUTS) allows us to automatically tune the number of steps L: we increases L until the simulated dynamics is long enough such that the proposed position θ^* starts to move back towards the initial position θ if we run more steps.

- This is measured by the angle between $\theta^* - \theta$ and current momentum ϕ^*.

A basic NUTS works as follows. Given the initial status,

1. Sample $u \mid \theta, \phi \sim \text{Uniform}\left[0, \exp\left\{-H\left(\theta, \phi\right)\right\}\right]$.
2. Apply the leapfrog method (with some modification) until a U-turn occurs.
3. Sample uniformly from the points in $\left\{(\theta, \phi) : \exp\left\{-H\left(\theta, \phi\right)\right\} \geq u\right\}$ that the leapfrog step has visited and the detailed balance condition is fulfilled.

Adaptively Tune ϵ

A too small ϵ will waste computation by taking needlessly tiny steps, and a too large will cause high rejection rates.

- In HMC, we tune ϵ in the warm-up stage of MCMC such that the average acceptance probability δ is the user specified value.
- In NUTS, there is no Metropolis accept/reject step. But we can still compute the ratio as if we were using the accept/reject step and set ϵ such that the pseudo acceptance probability is the user specified value.

In stan, the default is $\delta = 0.8$.

Burn-In Period

The stationary distribution is reached after large enough iterations.

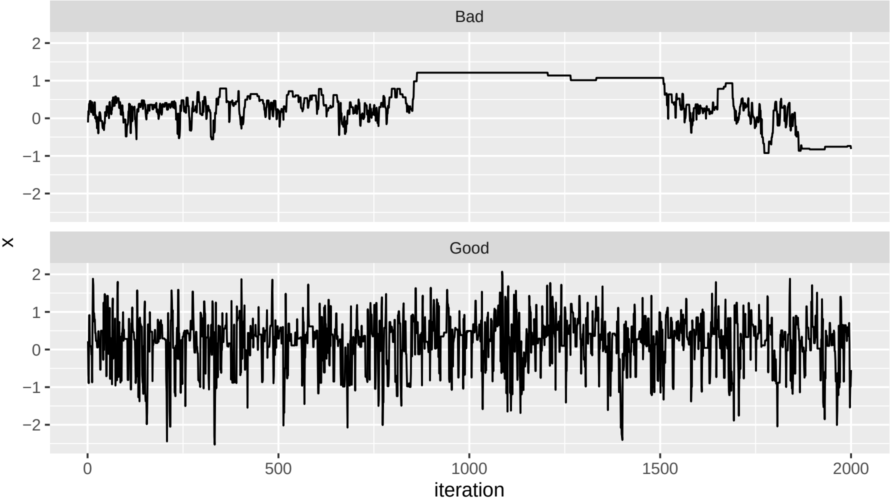
- If the iterations have not proceeded long enough, the simulated numbers may be unrepresentative of the target distribution.

To diminish the influence of the starting values, we can discard the early simulations, known as the burn-in.

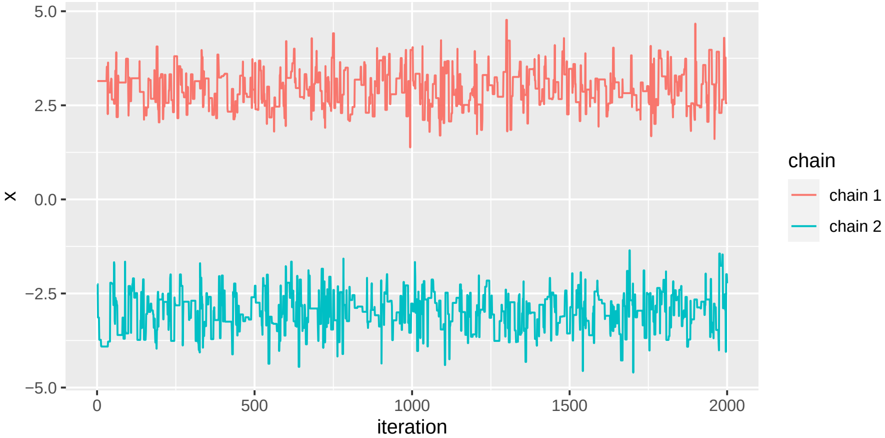
- There is no golden standard on how long the burn-in period should be.
- Hereafter, if the Markov chain has length n, we mean that after the burn-in period, the length is n.

Mixing

We want the Markov chain to show good mixing.



Several Markov Chains



One suggestion is to generate several independent Markov chains, starting from widely separated places.

Gelman-Rubin $\hat{R}$ Statistic: Variation

One way to assess convergence is the Gelman-Rubin $\hat{R}$ statistic. Suppose that we have simulated m chains each with n iterations. Say we have a univariate quantity $y_{ij} = f\left(\theta_j^{(i)}\right)$, where $\theta_j^{(i)}$ is the ith value in the jth chain.

- The variation within the chains is measured by

$$W = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{ij}-\bar{y}_{\cdot j}\right)^2\right],$$

  where $\bar{y}_{\cdot j}$ is the average of $\{y_{ij}\}_{i=1}^{n}$.
- The variation between the chains is measured by

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot}\right)^2,$$

  where $\bar{y}_{\cdot\cdot}$ is the average of all $\bar{y}_{\cdot j}$.

Gelman-Rubin $\hat{R}$ Statistic: Expression

If the Markov chains have reached stationary, then

$$\mathrm{E}\left[W\right] = \mathrm{E}\left[B\right] = \mathrm{Var}\left(Y\right).$$

- We estimate the variance $\mathrm{Var}\left(Y\right)$ by

$$\hat{V} = \frac{n-1}{n}W + \frac{1}{n}B.$$

- The Gelman-Rubin $\hat{R}$ statistic is then

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}},$$

  which declines to 1 as $n \to \infty$.

- It is suggested that we keep simulating the Markov chain until $\hat{R} < 1.1$ or even < 1.01.

Variants of Gelman-Rubin $\hat{R}$

Several different versions of $\hat{R}$ have been proposed.

1. One suggestion is to change $\hat{V}$ to

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} \left(1 + \frac{1}{m}\right) B.$$

   to account for the possibility that $\hat{V}$ is too low.

2. Another suggestion is to split each chain into two parts, yields $2m$ chains of length $n/2$ each. Then compute the $\hat{R}$, pretending that we have simulated $2m$ chains of length $n/2$.
   - This can be useful to detect the case where each chain does not reach stationary but the chains cover a common distribution, e.g, two chains exhibit an X-shape.

Serial Correlation

It is obvious that $\theta^{(t+1)}$ and $\theta^{(t)}$ are not independent draws. Inference from autocorrelated draws is generally less precise than from the same number of independent draws.

- However, such serial correlation is not necessarily a problem. Remember that, at convergence, we reach the stationary distribution.

```
Algorithm 4: General MCMC Integral
1 Sample a Markov chain for a given stationary distribution $\pi(\theta \mid x)$: $\theta^{(1)}$,
   ..., $\theta^{(R)}$ (after burn-in) ;
2 Approximate $\mu(x)$ by
      $\hat{\mu}^{\text{MCMC}} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i, x).$
```

Long-Run Property

Theorem

Under some conditions, for all starting state $\theta_0 \in \Theta$,

1. *ergodic theorem: For any initial state,*
$$\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) \quad \overset{a.s.}{\to} \quad E\left[h\left(\theta, x\right) \mid x\right] = \mu\left(x\right).$$
2. *central limit theorem: Let $\sigma^2 = Var\left[h\left(\theta, x\right) \mid x\right]$ and $\rho_j = corr\left(h\left(\theta^{(1)}, x\right), h\left(\theta^{(j+1)}, x\right) \mid x\right)$. Then,*
$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) - \mu\left(x\right)\right] \quad \overset{d}{\to} \quad N\left(0, \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

Effective Sample Size

If we have an iid sample of size n, then

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) - \mu\left(x\right)\right] \quad \overset{d}{\to} \quad N\left(0, \sigma^2\right).$$

If we have a converged Markov chain of length n,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) - \mu\left(x\right)\right] \quad \overset{d}{\to} \quad N\left(0,\ \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

The variance of $\hat{\mu}^{\text{MCMC}}$ is larger than the variance of $\hat{\mu}^{\text{IMC}}$. We define

$$n_{\text{eff}} \quad = \quad \frac{n}{1 + 2\sum_{j=1}^{\infty}\rho_j}$$

as the effective sample size of this Markov chain sample.

Estimate Effective Sample Size

We can also estimate the effective sample size, if we have m Markov chains of length n.

- Following the Gelman-Rubin $\hat{R}$ statistic, we can estimate σ^2 by

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} B.$$

- The autocorrelations can be estimated by

$$\hat{\rho}_t = 1 - \frac{\sum_{j=1}^{m} \sum_{i=t+1}^{n} \left(y_{i,j} - y_{i-t,j}\right)^2}{2m\left(n-t\right)\hat{V}}.$$

- The effective sample size is estimated by

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t},$$

  where T is the first odd positive integer such that $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative.

Alternative Confidence Interval

Suppose that we can divide the Markov chain of length n into b batches (e.g., 20 or proportional to $n^{1/3}$) of m consecutive observations each.

- Let $\bar{y}_j$ be the average of batch j.
- We will treat $\{\bar{y}_j\}$ as iid normal random variables.

An approximate confidence interval is

$$\frac{1}{b}\sum_{j=1}^{b}\bar{y}_j \pm t_{1-\alpha/2}\left(b-1\right)\sqrt{\frac{1}{b\left(b-1\right)}\sum_{j=1}^{b}\left(\bar{y}_j-\bar{y}\right)^2},$$

where $\bar{y}$ is the average of $\{\bar{y}_j\}$. This is just the usual t-confidence interval.

Thinning

Some prefer thinning the sequence by only keeping every kth draw from a sequence in order to reduce serial correlation.

- But whether or not the Markov chain is thinned, it can be used for inferences, provided that it has reached convergence.
- Suppose that the length of the Markov chain is n. We discard $k-1$ out of every k observations and the chain after thinning is n/k.
- Under some assumptions,

$$\begin{aligned} \sqrt{n}\left[\hat{\mu}-\mu\left(x\right)\right] &\overset{d}{\to} N\left(0,\ \tau^{2}\right), \\ \sqrt{n/k}\left[\hat{\mu}_k-\mu\left(x\right)\right] &\overset{d}{\to} N\left(0,\ \tau_k^{2}\right), \end{aligned}$$

  where $\hat{\mu}$ and $\hat{\mu}_k$ are the estimators without and with thinning, respectively.
- In fact, it has been proved that, for any $k>1$, $k\tau_k^2>\tau^2$, indicating that discarding $k-1$ out of every k observations will increase the variance.

Simulation Under Posterior

Using MCMC and other methods, we can simulate n random numbers from the posterior distribution $\pi(\theta \mid x)$. Using the simulated θ, we can

1. approximate the posterior mean: $n^{-1}\sum_{i=1}^{n}\theta^{(i)} \to \mathrm{E}[\theta \mid x]$.
2. approximate the posterior probability:
$$\frac{1}{n}\sum_{i=1}^{n} 1\left(\theta^{(i)} \in A\right) \quad\to\quad \mathrm{E}\left[1(\theta \in A) \mid x\right] = \mathrm{P}(\theta \in A \mid x).$$
3. approximate predictive density:
$$\frac{1}{n}\sum_{i=1}^{n} f\left(x_{\text{new}} \mid x, \theta^{(i)}\right) \quad\to\quad \int f(x_{\text{new}} \mid x, \theta)\,\pi(\theta \mid x)\,d\theta.$$
4. approximate mean of predictive distribution: $n^{-1}\sum_{i=1}^{n} x_{\text{new}}^{(i)} \to \mathrm{E}[x_{\text{new}} \mid x]$, where $x_{\text{new}}^{(i)}$ is simulated from $f\left(x_{\text{new}} \mid x, \theta^{(i)}\right)$.

Approximate Posterior

If the posterior distribution family is difficult to handle, it can be useful to approximate it by another distribution family that is easier to handle.

- The Kullback-Leibler divergence for distributions P and Q with respective densities p and q are

$$\mathrm{KL}\left(q, p\right) = \int q\left(\theta\right) \log\left[\frac{q\left(\theta\right)}{p\left(\theta\right)}\right] d\theta \geq 0.$$

- We choose a model $\mathcal{D}$ for the posterior, called the variational family.
- The variational density is

$$q^{*}\left(\theta \mid x\right) = \arg\min_{q \in \mathcal{D}} \mathrm{KL}\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right).$$

Variational Bayesian Inference

The idea of variational inference (VI) is to use $q^*(\theta \mid x) \in \mathcal{D}$ instead of $\pi(\theta \mid x)$ and to explore the properties of $\mathcal{D}$.

We need to choose $\mathcal{D}$ ourselves.

- Trade-off: too simple $\mathcal{D}$ poorly approximates $\pi(\theta \mid x)$ but too complex $\mathcal{D}$ is hard to handle.
- One choice is the mean-field variational family $\mathcal{D}_{\text{MF}}$, where

  $$q(\theta \mid x) = \prod_{j=1}^{m} q_j(\theta_j \mid x),$$

  that is, the components in θ are independent. We call $q_j(\theta_j \mid x)$ the jth variational factor.

Evidence Lower Bound

The Kullback-Leilber divergence satisfies

$$\mathrm{KL}\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right) = \log\left[m\left(x\right)\right] - \underbrace{\int q\left(\theta \mid x\right)\log\left[\frac{p\left(\theta, x\right)}{q\left(\theta \mid x\right)}\right] d\theta}_{\text{evidence lower bound ELBO}(q)}.$$

Since $\mathrm{KL}\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right) \geq 0$, the ELBO satisfies

$$\mathrm{ELBO}\left(q\right) \leq \log\left[m\left(x\right)\right],$$

a lower bound of the log-marginal likelihood of x.

- Minimizing the KL divergence is the same as maximization of ELBO.

Variational Inference in Linear Regression

Example

Suppose that $y \mid \beta \sim N_n\left(X\beta, \Sigma\right)$ and $\beta \sim N_p\left(\mu_0, \Lambda_0^{-1}\right)$, where Σ is known. The posterior is $\beta \mid y \sim N\left(\mu_n, \Lambda_n^{-1}\right)$, where

$$\begin{aligned}\Lambda_n &= \Lambda_0 + X^T\Sigma^{-1}X, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0\mu_0 + X^T\Sigma^{-1}y\right).\end{aligned}$$

Consider the mean-field variational family

$$\mathcal{D}_{\mathrm{MF}} = \left\{N_p\left(\mu, \Lambda^{-1}\right) : \ \mu \in \mathbb{R}^p, \Lambda \text{ is diagonal}\right\}.$$

Find the ELBO and the variational density.

Explicit Expression of $\mathcal{D}_{\mathrm{MF}}$

Theorem

Consider the mean field variational family $\mathcal{D}_{MF}$, where

$$q\left(\theta \mid x\right) \;=\; \prod_{j=1}^{m} q_j\left(\theta_j \mid x\right).$$

Let θ_k be the kth group in θ and

$$q_k^*\left(\theta_k \mid x\right) \;=\; \arg\min_{q_k} KL\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right).$$

Then,

$$q_k\left(\theta_k \mid x\right) \;\propto\; \exp\left\{\int q_{-k}\left(\theta_{-k} \mid x\right)\log\pi\left(\theta_k \mid \theta_{-k}, x\right)d\theta_{-k}\right\}.$$

Coordinate Ascent Variational Inference Algorithm

The previous theorem suggests the following stepwise conditioning to approximate $q^* (\theta \mid x)$.

Algorithm 5: Coordinate ascent variational inference (CAVI) Algorithm

```
Choose an initial approximation $\hat{q}^{(0)} (\theta \mid x) = \prod_{j=1}^{m} \hat{q}_j^{(0)} (\theta_j \mid x)$ ;
for $t = 1$ in 1 : $T$ do
    for $j = 1$ in 1 : $m$ do
        Calculate $\hat{q}_j^{(t)} (\theta_j \mid x) \propto \exp \left\{ \int q_{-j} (\theta \mid x) \log \pi (\theta_j \mid \theta_{-j}, x) \, d\theta_{-j} \right\}$,
          where
          $$q_{-j} (\theta \mid x) = \left[ \prod_{k=1}^{j-1} \hat{q}_k^{(t)} (\theta_k \mid x) \right] \left[ \prod_{k=j+1}^{m} \hat{q}_k^{(t-1)} (\theta_k \mid x) \right]$$
    end
end
```

CAVI Algorithm: Example

Example

Suppose that we have an iid sample $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$. The priors are $\mu \mid \sigma^2 \sim N\left(\mu_0, \sigma^2/\lambda_0\right)$ and $\sigma^2 \sim \text{InvGamma}\left(a_0, b_0\right)$. The posterior is

$$\mu \mid x, \sigma^2 \sim N\left(\mu_n,\ \sigma^2/\lambda_n\right), \quad \text{and} \quad \sigma^2 \mid x \sim \text{InvGamma}\left(a_n, b_n\right).$$

where $\lambda_n = \lambda_0 + n$, $\mu_n = \lambda_n^{-1}\left(\lambda_0\mu_0 + \sum_{i=1}^n x_i\right)$, $a_n = a_0 + \frac{n}{2}$, and

$$b_n = b_0 + \frac{1}{2}\left(\sum_{i=1}^n x_i^2 + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right).$$

Stan

Stan is a c++ library for Bayesian inference using HMC to obtain posterior simulations.

- Rstan is the R interface to Stan.
- PyStan is the Python interface to Stan.

It is the state-of-the-art library for doing Bayesian statistics.

A Stan model consists of

1. data,
2. parameters,
3. statistical model.

R Package rstanarm

The R package rstanarm emulates the R syntax but uses Stan via the rstan package to fit models in the background. So you skip writing the Stan syntax.

- Various common regression models have been implemented in rstanarm.
- Another benefit is that various visualization tools in R can be used.

Bayesian Statistics
Chapter 7: Model Comparison

Shaobo Jin

Department of Mathematics

Candidate Models

Suppose that we have a set of candidate models

$$\mathcal{M}_k : \ x \sim f_k \left(x \mid \theta_k \right), \qquad \theta_k \in \Theta_k,$$

where $\{f_k\}$ may belong to the same distribution family but have different parameter space, or $\{f_k\}$ may belong to different distribution families.

Example

$$\begin{aligned} H_0 : & \qquad Y \sim N\left(X_1\beta_1, \sigma^2\right) \text{ with conjugate prior} \\ H_1 : & \qquad Y \sim N\left(X_1\beta_1 + X_2\beta_2, \sigma^2\right), \text{ with conjugate prior.} \end{aligned}$$

Or

$$\begin{aligned} H_0 : & \qquad X \sim \text{Binomial}\left(n, p\right), \ p \sim \text{Beta}\left(a_0, b_0\right) \\ H_1 : & \qquad X \sim \text{Poisson}\left(\lambda\right), \ \lambda \sim \text{Gamma}\left(a_1, b_1\right). \end{aligned}$$

Hierarchical Prior

When we compare several candidate models, the model index is also viewed as a parameter. The prior allocation is hierarchical:

1. prior probability for model $\mathcal{M}_k$: $p_k = \mathrm{P}\left(\mathcal{M}_k \text{ is the true model}\right)$,
2. given model $\mathcal{M}_k$, prior $\pi_k\left(\theta_k\right)$ for parameter θ_k in $\mathcal{M}_k$.

The posterior of interest is now

$$\begin{aligned} \mathrm{P}\left(\mathcal{M}_k \mid x\right) &= \frac{p_k \int_{\Theta_k} f_k\left(x \mid \theta_k\right) \pi_k\left(\theta_k\right) d\theta_k}{\sum_j p_j \int_{\Theta_j} f_j\left(x \mid \theta_j\right) \pi_j\left(\theta_j\right) d\theta_j} \\ &\propto \mathrm{P}\left(\mathcal{M}_k\right) f\left(x \mid \mathcal{M}_k\right). \end{aligned}$$

Bayes Factor

We have used the Bayes factor in hypothesis testing. It can also be used in model comparison.

Definition

For two Bayes models $\mathcal{M}_1$ and $\mathcal{M}_2$, the Bayes factor is

$$\begin{aligned} B_{12} &= \frac{\mathrm{P}\left(\mathcal{M}_1 \mid x\right)/\mathrm{P}\left(\mathcal{M}_2 \mid x\right)}{\mathrm{P}\left(\mathcal{M}_1\right)/\mathrm{P}\left(\mathcal{M}_2\right)} \\ &= \frac{\int_{\Theta_1} f_1\left(x \mid \theta_1\right)\pi_1\left(\theta_1\right)d\theta_1}{\int_{\Theta_2} f_2\left(x \mid \theta_2\right)\pi_2\left(\theta_2\right)d\theta_2}, \end{aligned}$$

where

$$m_k\left(x\right) = \int\limits_{\Theta_k} f_k\left(x \mid \theta_k\right)\pi_k\left(\theta_k\right)d\theta_k$$

is the marginal likelihood under model k.

Compute Bayes Factor: Example

It is super important to keep in mind that we cannot use $\propto$ anymore. We must keep track of all normalizing constants!

Example

Suppose that we have randomly chosen n patients and analyzed their blood samples in order to test drug resistance. Let X be the number of patients with positive test result. Two models are under consideration

$$\begin{aligned} \mathcal{M}_1: &\quad X \sim \text{Binomial}(n, p), \; p \sim \text{Beta}(a_0, b_0) \\ \mathcal{M}_2: &\quad X \sim \text{Poisson}(\lambda), \; \lambda \sim \text{Gamma}(a_1, b_1). \end{aligned}$$

Compute the Bayes factor.

Bayes Factor for Nested Linear Model

Suppose that we want to compare two nested linear regression models

$$\begin{aligned}
\mathcal{M}_1: &\quad Y \sim N_n\left(X_1\beta_1, \sigma^2 I_n\right),\ \left(\beta_1, \sigma^2\right) \sim \text{NIG}\left(a_0, b_0, \tau_0, \Omega_0^{-1}\right),\\
\mathcal{M}_2: &\quad Y \sim N_n\left(X\beta, \sigma^2 I_n\right),\ \left(\beta, \sigma^2\right) \sim \text{NIG}\left(a_0, b_0, \mu_0, \Lambda_0^{-1}\right),
\end{aligned}$$

where $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ such that $X\beta = X_1\beta_1 + X_2\beta_2$.

- Comparing $\mathcal{M}_1$ and $\mathcal{M}_2$ is the same as testing $\beta_2 = 0$.

Example

Compute the Bayes factor for the above nested linear regression models.

Reminder

Keep in mind that using the Bayes factor to conduct model selection and hypothesis testing share many things in common.

- We need to be careful when using improper prior with Bayes factors, as the Bayes factor can be biased towards certain model.
- The Jeffreys-Lindley paradox says that an improper prior cannot be approximated by priors with increasing variances.

Bayesian Information Criterion

Using the marginal likelihood under model k, we call

$$2\log m_k(x) = 2\log\left[\int_{\Theta_k} f_k(x \mid \theta_k)\,\pi_k(\theta_k)\,d\theta_k\right]$$

the exact Bayesian information criterion (BIC) values.

- However, it is rarely computed in practice due to the complexity of the integral.

Definition

The Bayesian information criterion (BIC), aka, Schwartz's criterion, is

$$\mathrm{BIC} = -2\max_{\theta}\log f(x \mid \theta) + p\log(n),$$

where n is the sample size and p is the dimension of θ.

Derivation of BIC

We approximate $m_k(x)$ by the Laplace approximation and obtain

$$m_k(x) \approx (2\pi)^{p_k/2} \sqrt{\det\left(\left[\frac{\partial^2 - \log f_k\left(x \mid \hat{\theta}_k\right)}{\partial\theta_k \partial\theta_k^T}\right]^{-1}\right)} f\left(x \mid \hat{\theta}_k\right) \pi_k\left(\hat{\theta}_k\right)$$

where $\hat{\theta}_k$ is the MLE under $\mathcal{M}_k$ such that $\frac{\partial \log f_k\left(x|\hat{\theta}_k\right)}{d\theta_k} = 0$. Hence,

$$-2\log m_k(x) \approx \underbrace{-2\log f\left(x \mid \hat{\theta}_k\right) + p_k \log n}_{\text{BIC}_k} + O_P(1).$$

It is interesting to see that the prior vanishes in BIC.

BIC: Example

Example

Suppose that $Y \mid \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right)$ and $\pi\left(\beta, \sigma^2\right) = \sigma^{-2}$. Find the BIC.

Implications of Approximation

1. Since the Bayes factor is $B_{12} = \frac{m_1(x)}{m_2(x)}$, then

$$2\log B_{12} \quad \approx \quad \mathrm{BIC}_2 - \mathrm{BIC}_1.$$

   We favor $\mathcal{M}_2$ if B_{12} is small, or equivalently $m_2(x)$ is large. This is similar to favoring $\mathcal{M}$ if BIC_2 is small.

2. BIC penalizes the log-likelihood with $p\log n$.
   - For nested models $\mathcal{M}_1 \subset \mathcal{M}_2$ such that $\theta_1 \subset \theta_2$,

$$\mathrm{P}_{\mathcal{M}_1}\left(-2\log\left[\frac{f\left(x \mid \hat{\theta}_1\right)}{f\left(x \mid \hat{\theta}_2\right)}\right] < c \mid \theta_1\right) \quad \to \quad \mathrm{P}_{\mathcal{M}_1}\left(\chi^2_{p_2-p_1} < c \mid \theta_1\right),$$

     as $n \to \infty$. Without the penalization term, the probability of choosing the correct model is not 1.

Effect of Penalization

Consider the information criterion of the form

$$\mathrm{IC}_k = -2\sum_{i=1}^{n} \log f_k\left(x_i \mid \hat{\theta}_k\right) + c_k,$$

where $c_k > 0$. We select the candidate model that has the smallest information criterion.

Theorem (Weakly Consistency)

Suppose that there is only one candidate model that minimizes the Kullback-Leibler divergence to the true model. If $c_k = o_P(n)$, then the information criterion selects such closest model with probability approaching 1 as $n \to \infty$.

Both Akaike information criterion (AIC, $c_k = 2p_k$) and BIC ($c_k = p \log n$) are weakly consistent.

Extent of Penalization

Theorem (Consistency)

Suppose that there are several candidate models that minimize the Kullback-Leibler divergence. Let $\mathcal{J}$ be the set of indices of candidate models which all reach the minimum Kullback-Leibler divergence to the true model, and $\mathcal{J}_0$ be the subset of $\mathcal{J}$ containing the smallest dimensions. Suppose that, for any $j_0 \in \mathcal{J}_0$ and $j \in \mathcal{J} \setminus \mathcal{J}_0$, we have

$$P\left(c_j - c_{j_0} \to \infty,\ \text{as } n \to \infty\right) \ = \ 1.$$

Then the information criterion selects the most parsimonious model from the closest models with probability approaching 1.

AIC ($c_k = 2p_k$) does not fulfill the condition of the theorem, whereas BIC ($c_k = p \log n$) is consistent.

Deviance

Definition

The deviance for a given model and given data x is $D(\theta) = -2\log f(x \mid \theta) + \text{constant}$, where the constant is the same in all candidate models, representing the log-likelihood of the perfectly fitted model.

- The difference in the deviance can be used to compare the fit of candidate models to the data such as

$$D\left(\hat{\theta}_1\right) - D\left(\hat{\theta}_2\right) = 2\left[\log f\left(x \mid \hat{\theta}_2\right) - \log f\left(x \mid \hat{\theta}_1\right)\right].$$

- A Bayesian version such deviance difference is

$$D(\theta) - D(\mathrm{E}[\theta \mid x]) = 2\left[\log f(x \mid \mathrm{E}[\theta \mid x]) - \log f(x \mid \theta)\right].$$

Deviance Information Criterion

Let

$$\begin{aligned} p_D &= \mathrm{E}\left[D\left(\theta\right) - D\left(\mathrm{E}\left[\theta \mid x\right]\right) \mid x\right] \\ &= \mathrm{E}\left[D\left(\theta\right) \mid x\right] - D\left(\mathrm{E}\left[\theta \mid x\right]\right) \end{aligned}$$

be the posterior expected value of the deviance difference.

Definition

The deviance information criterion (DIC) is

$$\begin{aligned} \mathrm{DIC} &= \mathrm{E}\left[D\left(\theta\right) \mid x\right] + p_D \\ &= D\left(\mathrm{E}\left[\theta \mid x\right]\right) + 2p_D \\ &= -4\mathrm{E}\left[\log f\left(x \mid \theta\right) \mid x\right] + 2\log f\left(x \mid \mathrm{E}\left[\theta \mid x\right]\right) + \text{constant}. \end{aligned}$$

We choose the model with a smaller value of DIC.

Example: DIC for Linear Model

Example

Consider the linear regression model

$$Y \sim N_n\left(X\beta, \sigma^2 I_n\right), \qquad \left(\beta, \sigma^2\right) \sim \mathrm{NIG}\left(a_0, b_0, \mu_0, \Lambda_0^{-1}\right).$$

Compute the DIC. It is known that, if $\sigma^2 \sim \mathrm{InvGamma}\,(a, b)$, then

$$\begin{aligned} \mathrm{E}\left[\sigma^2\right] &= \frac{b}{a-1}, \text{ if } a > 1 \\ \mathrm{E}\left[\log \sigma^2\right] &= \log\left(b\right) - \psi\left(a\right), \\ \mathrm{E}\left[\sigma^{-2}\right] &= \frac{a}{b}. \end{aligned}$$

where ψ is the digamma function.

DIC and AIC

Suppose that posterior can be well approximated by a normal distribution

$$\theta \mid x \;\; \approx \;\; N_p\left(\hat{\theta},\; \left[-\frac{\partial^2 \log f\left(x \mid \hat{\theta}\right)}{\partial\theta\partial\theta^T}\right]^{-1}\right),$$

where $\hat{\theta}$ is the MLE. Then, we can approximate $\mathrm{E}\left[\theta \mid x\right]$ by $\hat{\theta}$.

- We will investigate such normal approximation in a later lecture.

We can show that

$$\mathrm{DIC} \;\; \approx \;\; -2\log f\left(x \mid \hat{\theta}\right) + 2p + \mathrm{constant}.$$

which is equivalent to AIC.

Computational Technique

To compute DIC, we need to evaluate the integrals

$$\begin{aligned} \mathrm{E}\left[D\left(\theta\right) \mid x\right] &= \int D\left(\theta\right) \pi\left(\theta \mid x\right) d\theta, \\ D\left(\mathrm{E}\left[\theta \mid x\right]\right) &= D\left(\int \theta \pi\left(\theta \mid x\right) d\theta\right). \end{aligned}$$

The integrals are generally intractable and need to be evaluated numerically.

We can for example draw posterior samples θ_1, ..., θ_T by independent MC or MCMC and approximate the integrals by

$$\begin{aligned} \int D\left(\theta\right) \pi\left(\theta \mid x\right) d\theta &\approx \frac{1}{n} \sum_{t=1}^{T} D\left(\theta_i\right), \\ \int \theta \pi\left(\theta \mid x\right) d\theta &\approx \frac{1}{n} \sum_{t=1}^{T} \theta_i. \end{aligned}$$

Combining Different Models

A different view relative to model selection is to combine the contributions of several models as in ensemble learning. Let Δ be the quantity of interest such as

- average treatment effect of a drug,
- a future value.

Bayesian model selection chooses a model k^* using $\mathrm{P}\left(\mathcal{M}_k \mid x\right)$ and estimates Δ by

$$\hat{\Delta} = \mathrm{E}\left[\Delta \mid x, \mathcal{M}_{k^*}\right].$$

Bayesian model averaging (BMA) takes a weighted average instead as

$$\hat{\Delta} = \sum_k \mathrm{E}\left[\Delta \mid x, \mathcal{M}_k\right] \mathrm{P}\left(\mathcal{M}_k \mid x\right).$$

Posterior of Δ

The posterior of Δ is given by

$$f\left(\Delta \mid x\right) = \sum_{k} f\left(\Delta \mid \mathcal{M}_k, x\right) \mathrm{P}\left(\mathcal{M}_k \mid x\right),$$

where $f\left(\Delta \mid \mathcal{M}_k, x\right)$ is the posterior of Δ under model k. The posterior mean of Δ is

$$\begin{aligned} \mathrm{E}\left[\Delta \mid x\right] &= \int \Delta \sum_{k} f\left(\Delta \mid \mathcal{M}_k, x\right) \mathrm{P}\left(\mathcal{M}_k \mid x\right) d\Delta \\ &= \sum_{k} \mathrm{P}\left(\mathcal{M}_k \mid x\right) \underbrace{\int \Delta f\left(\Delta \mid \mathcal{M}_k, x\right) d\Delta}_{=\mathrm{E}\left[\Delta \mid x, \mathcal{M}_k\right]}, \end{aligned}$$

which is the BMA estimator of Δ.

Three Scenarios

The posterior probability

$$\mathrm{P}\left(\mathcal{M}_k \mid x\right) \;=\; \mathrm{P}\left(\mathcal{M}_k \text{ is the true model} \mid x\right)$$

can be strange if all candidate models are wrong, especially when we need to specify the prior probability that $\mathcal{M}_k$ is the true model.

1. The $\mathcal{M}$-closed setting means that one of the candidate models is the true data generating process.
2. The $\mathcal{M}$-complete setting means that but the true data generating process can be conceptualized, but it is not one of the candidate models due to, for example, model complexity or lack of information.
3. The $\mathcal{M}$-open setting means that the data generating process cannot be conceptualized and all candidate models are wrong.

Bayesian Stacking

Let $S(P, Q)$ be a scoring rule to measure the similarity between two probability measure P and Q. Let p and q be the corresponding densities. Then,

$$S(P, Q) = \int s(P, w)\, q(w)\, dw,$$

for some function $s(\cdot, \cdot)$. Bayesian stacking maximizes such similarity

$$S\left(\sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k),\ f_{\text{true}}(\tilde{x} \mid x)\right)$$

with respect to weights $\{w_k\}$ under the restriction that

$$\sum_k w_k = 1,\ 0 \leq 1 w_k \leq 1,\ \forall k.$$

Scoring Rule

Two commonly used scoring rules are

1. log score: $s(P, x) = \log p(x)$ such that $S(Q, Q) - S(P, Q)$ is the KL divergence.
   - Taking $p(x) = \sum_k w_k f(\tilde{x} \mid x, \mathcal{M}_k)$ is the same as maximizing the similarity between the stacked predictive distribution and the true predictive distribution.
2. energy score: $s(P, x) = \frac{1}{2}\mathrm{E}_P\left[\left\|X - \tilde{X}\right\|^\beta\right] - \mathrm{E}_P\left[\|X - x\|^\beta\right]$, where X and $\tilde{X}$ are two iid random variables, and the expectations are taken with respect to P.
   - If $\beta = 2$, it reduces to $s(P, x) = -\|\mathrm{E}_P[X] - x\|^2$.
   - Maximizing the scoring rule is equivalent to minimizing the squared prediction error.

Leave-One-Out Cross Validation

However, we don't know $f_{\text{true}}\left(\tilde{x} \mid x\right)$ that is needed to evaluate

$$S\left(\sum_k w_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right),\ f_{\text{true}}\left(\tilde{x} \mid x\right)\right).$$

One alternative is to use leave-one-out cross validation as

$$\min_w \frac{1}{n}\sum_{i=1}^{n} s\left(\sum_k w_k f\left(x_i \mid x_{-i}, \mathcal{M}_k\right),\ x_i\right),$$

where

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int f\left(x_i \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k.$$

The stacked estimate of the predictive density is

$$\hat{f}\left(\tilde{x} \mid x\right) = \sum_k \hat{w}_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right).$$

BMA and Bayesian Stacking

For BMA, it is alleged that, as $n \to \infty$,

- if one of the candidate models is the true model, say $\mathcal{M}_{k^*}$ is the true model,
- or, if all candidate models are misspecified and $\mathcal{M}_{k^*}$ has the smallest Kullback-Leibler divergence to the true model,

then $\mathrm{P}\left(\mathcal{M}_{k^*} \mid x\right) \to 1$.

In contrast to Bayesian model averaging,

- no prior $\mathrm{P}\left(\mathcal{M}_k\right)$ is needed in Bayesian stacking.
- Bayesian stacking is intended for the case where all candidate models are misspecified.

Importance Sampling

It is computationally intensive to compute the leave-one-out (LOO) predictive density

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k$$

for each i, because we have to refit $\mathcal{M}_k$ n times to obtain all $\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)$.

Suppose that, for each k, we fit $\mathcal{M}_k$ using all the data and obtain L draws from the posterior $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$. Then,

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \pi\left(\theta_k \mid x, \mathcal{M}_k\right) d\theta_k,$$

where the importance weight is

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)}.$$

Normalized Importance Sampling

We can rewrite the importance weight as

$$\begin{aligned} w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} & \propto \frac{f\left(x_{-i} \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)}{f\left(x \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid \mathcal{M}_k\right)} \\ & \propto \frac{1}{f\left(x_i \mid \theta_k, \mathcal{M}_k\right)}. \end{aligned}$$

The normalized importance sampling estimator is

$$\hat{f}^{\text{NIS}}\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \frac{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right) p\left(x_i \mid \theta_k, \mathcal{M}_k\right)}{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)},$$

where we sample $\theta_k^{(l)}$, $l = 1, ..., L$, from $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$.

However, the importance weights can be unstable if the distribution has a long tail that makes the importance weight very large.

Generalized Pareto Distribution

Theorem

Under suitable conditions on the random variable X, if the threshold u_0 is high enough, the conditional distribution of $X \mid X > u$ converges to a three-parameter generalized Pareto distribution (GPD), as $u \to \infty$. Its density is given by

$$f\left(x; u, \sigma, k\right) = \begin{cases} \frac{1}{\sigma}\left[1+\frac{k(x-u)}{\sigma}\right]^{-1-1/k}, & \text{if } c \neq 0, \\ \frac{1}{\sigma}\exp\left(\frac{x-u}{\sigma}\right), & \text{if } c = 0, \end{cases}$$

for $y > u$ and $\sigma > 0$.

Pareto Smoothed Importance Sampling

Pareto Smoothed Importance Sampling stabilizes the large importance weights. Without loss of generality, suppose that $\left\{w_i\left(\theta_k^{(l)},\mathcal{M}_k\right)\right\}$ has been ordered in increasing order.

- Consider the largest $N=\left\lfloor\min\left(0.2L,\,3\sqrt{L}\right)\right\rfloor$ importance weights.
- We fit a GPD to $\left(w_i\left(\theta_k^{(L-N+1)},\mathcal{M}_k\right),...,w_i\left(\theta_k^{(L)},\mathcal{M}_k\right)\right)$ with $u=w_i\left(\theta_k^{(L-N)},\mathcal{M}_k\right)$.
- These N tail importance weights are replaced by

$$\min\left\{F^{-1}\left(\frac{z-1/2}{M}\right),\ \max_i w_i\left(\theta_k^{(l)},\mathcal{M}_k\right)\right\},\qquad z=1,...,M,$$

  where F^{-1} is the inverse distribution function of the fitted GPD. The other importance weights are unchanged.

Bayesian Statistics
Hierarchical Model and Empirical Bayes

Shaobo Jin

Department of Mathematics

Hierarchical Model

It is often the case that we can easily define a probabilistic model through several levels of conditional distribution, instead of a joint distribution.

- One trivial example is our Bayesian analysis: we specify $f(x \mid \theta)$ and $\pi(\theta)$, instead of $f(x, \theta)$.

We can further introduce hierarchical modeling to

- $f(x \mid \theta)$ to model complex structure in x.
- $\pi(\theta)$ as hierarchical prior.

Incomplete Prior

It is often the case that we have some prior information, but it is not enough for us to fully determine the prior.

Example

For example, we want to specify a beta prior for success probability. Our prior information suggests that such probability is around 0.3.

We need to take the uncertainty in specifying the prior into account in our Bayesian modeling such as

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

More levels in the prior can be introduced similarly.

Hierarchical Prior: Uncertainty in Prior

Example

Suppose that $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. We want to use a beta prior $\theta \sim \text{Beta}\left(a_0, b_0\right)$. Instead of specifying the values of a_0 and b_0 such that $a_0 / \left(a_0 + b_0\right) = 0.3$ directly, we reparametrize the beta density into

$$\pi\left(\theta\right) = \frac{1}{B\left(\mu\kappa, \left(1-\mu\right)\kappa\right)} \theta^{\mu\kappa - 1} \left(1-\theta\right)^{\left(1-\mu\right)\kappa - 1},$$

such that $a_0 = \mu\kappa$ and $b_0 = \left(1-\mu\right)\kappa$.

- Specifying the values of a_0 and b_0 is equivalent to specifying the values of μ_0 and κ_0 directly.
- Instead, we introduce a prior $\pi\left(\mu, \kappa\right)$ to μ and κ.

Prior Distribution

Suppose that

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

- If λ is known (e.g., we specify its value direct), then $\pi(\theta \mid \lambda)$ is the usual prior. If λ is unknown, then $\pi(\lambda)$ is a prior on a prior (i.e., hyperprior).
- The prior $\pi(\theta)$ is obtained by

$$\pi(\theta) = \int \pi(\theta \mid \lambda) \pi(\lambda) d\lambda.$$

Hierarchical Prior: Another Example

A general characteristic of the hierarchical prior is that it introduces robustness.

Example

Suppose that $Y \mid \theta \sim N\left(\theta,\, \sigma^2\right)$, where σ^2 is known. The conjugate prior is $\theta \sim N\left(\mu_0, \lambda_0^{-1}\right)$. Instead of specifying the prior for θ directly, we consider the hierarchical prior

$$\begin{aligned} \theta \mid \tau^2 &\sim N\left(\mu_0, \tau^2\right), \\ \tau^2 &\sim \operatorname{InvGamma}\left(a_0, b_0\right). \end{aligned}$$

This hierarchical prior turns out to be equivalent to specifying a t distribution prior on θ, which has a heavier tail than a normal prior.

Hierarchical Prior: One More Example

Example

Suppose that we want to model data from different countries

$$Y_{ij} \mid \theta_j \quad \sim \quad N\left(\theta_j, \sigma^2\right),$$

where Y_{ij} is the ith observation from the jth country. It is often natural to assume that data from each country is draw from a bigger population, and we assume

$$\theta_j \mid \mu \quad \sim \quad N\left(\mu, \omega^2\right).$$

But we don't know μ, so we specify a prior for it $\mu \sim \pi\left(\mu\right)$. This idea is Bayesian meta analysis that allows us to combine similar studies conducted by different entities.

Full Bayesian Treatment

Consider again

$$\begin{aligned} x \mid \theta &\sim f(x \mid \theta), \\ \theta \mid \lambda &\sim \pi(\theta \mid \lambda), \\ \lambda &\sim \pi(\lambda). \end{aligned}$$

Neither θ nor λ is known. A full Bayesian treatment of such hierarchical model is based on the joint prior

$$\pi(\theta, \lambda) = \pi(\theta \mid \lambda) \pi(\lambda)$$

and the joint posterior

$$\pi(\theta, \lambda \mid x) \propto f(y \mid \theta, \lambda) \pi(\theta, \lambda) = f(y \mid \theta) \pi(\theta \mid \lambda) \pi(\lambda).$$

For example, we can use MCMC to draw posterior samples from $\pi(\theta, \lambda \mid x)$.

Hierarchical Prior: Example

Example

Suppose that $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$. Consider the hierarchical prior

$$\begin{aligned} \theta \mid \mu, \kappa &\sim \text{Beta}\left(\mu\kappa,\ \left(1-\mu\right)\kappa\right), \\ \mu &\sim \text{Uniform}\left[0,1\right], \\ \kappa &\sim \text{Exp}\left(1\right), \end{aligned}$$

where the priors of μ and κ are independent. Find the posterior distributions.

Empirical Bayes

Instead of directly introducing a hyperprior on θ and apply the full Bayesian treatment, the empirical Bayes approach estimates the unknown hyperparameters from the marginal distribution and use the Bayes formula treating $\hat{\pi}$ as a prior.

- Suppose that $x \mid \theta \sim f\left(x \mid \theta\right)$ and $\theta \mid \lambda \sim \pi\left(\theta \mid \lambda\right)$, where λ is unknown.
- If λ were known, the posterior is the usual

$$\pi\left(\theta \mid x\right) \quad \propto \quad f\left(x \mid \theta\right)\pi\left(\theta \mid \lambda\right).$$

- The empirical Bayes approach estimates λ from $f\left(x \mid \lambda\right)$, and apply the "plug-in principle" by treating $\pi\left(\theta \mid \hat{\lambda}\right)$ as the prior of θ.

Empirical Bayes: Example

Example

Find a point estimator of θ using empirical Bayes.

1. Suppose that we observe independent $X_i \mid \theta_i \sim \text{Binomial}\left(m, \theta_i\right)$, $i = 1, ..., n$. Consider the independent prior $\theta_i \mid a, b \sim \text{Beta}\left(a,\ b\right)$. The posterior is $\theta_i \mid x_i \sim \text{Beta}\left(a + x_i, b + m - x_i\right)$. If a and b are known, the posterior mean is

$$\text{E}\left[\theta \mid x, a, b\right] = \frac{a + x}{a + b + m}.$$

2. Suppose that $x \mid \theta \sim N_p\left(\theta, I_p\right)$ and the prior distributions ofθ_i are independent $N\left(0, \tau^2\right)$. If τ^2 is known, the posterior is

$$\pi\left(\theta \mid x\right) \sim N\left(\frac{\tau^2}{\tau^2 + 1} x,\ \frac{\tau^2}{\tau^2 + 1} I_p\right).$$

Hierarchical Data

The hierarchical structure is not only limited to prior. We can also use the conditional distributions to help us specify the likelihood part.

- A natural example is the Gaussian mixture model, where

$$f\left(x \mid \theta\right) \quad = \quad pN\left(\mu_1, \sigma_1^2\right) + \left(1-p\right) N\left(\mu_2, \sigma_2^2\right).$$

- This likelihood is generally difficult to handle. But we can introduce a latent variable $Z \sim \text{Bernoulli}\left(p\right)$ and specify the likelihood as

$$\begin{aligned} f\left(x \mid z, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) &= \left[N\left(\mu_1, \sigma_1^2\right)\right]^z \left[N\left(\mu_2, \sigma_2^2\right)\right]^{1-z}, \\ f\left(z \mid p\right) &= p^z \left(1-p\right)^{1-z}. \end{aligned}$$

Gaussian Mixture Model

Consider the Gaussian mixture model

$$f\left(x \mid \theta\right) \;\; = \;\; pN\left(\mu_1, \sigma_1^2\right) + \left(1-p\right)N\left(\mu_2, \sigma_1^2\right).$$

In a full Bayesian treatment, we can develop a Gibbs sampler to sample from the posterior

$$f\left(z, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p \mid x\right).$$

However, MCMC will encounter various problems.

- Label switching: for simplicity let $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the likelihood of $\left(p, \mu_1, \mu_2, \sigma^2\right) = \left(0.2, 0, -1, \sigma^2\right)$ will produce the same likelihood as $\left(p, \mu_1, \mu_2, \sigma^2\right) = \left(0.8, -1, 0, \sigma^2\right)$.
- Component collapsing: a mixture of K normals can yield $\mu_i = \mu_j$ and $\sigma_i = \sigma_j$ even $i \neq j$.

Latent Variable Model

The finite mixture model is a special case of a latent variable model, where the mixture indicators $\{z_i\}$ are latent. Another example of a latent variable model is the measurement error model.

- Suppose that we want to observe $z \sim f(z \mid \lambda)$, but we only observe $x = z + e$, where e is the measurement error.
- If z observed, then the usual Bayes analysis yields

$$\pi(\lambda \mid z) \quad \propto \quad f(z \mid \lambda)\,\pi(\lambda).$$

- Since only x is observed, we need to use a hierarchical model

$$\pi(z, \lambda \mid x) \quad \propto \quad f(x \mid z)\,f(z \mid \lambda)\,\pi(\lambda),$$

  where we have treated z as nuisance parameters.

Combining Different Models

A different view relative to model selection is to combine the contributions of several models as in ensemble learning. Let Δ be the quantity of interest such as

- average treatment effect of a drug,
- a future value.

Bayesian model selection chooses a model k^* using $\mathrm{P}\left(\mathcal{M}_k \mid x\right)$ and estimates Δ by

$$\hat{\Delta} \ = \ \mathrm{E}\left[\Delta \mid x, \mathcal{M}_{k^*}\right].$$

Bayesian model averaging (BMA) takes a weighted average instead as

$$\hat{\Delta} \ = \ \sum_k \mathrm{E}\left[\Delta \mid x, \mathcal{M}_k\right] \mathrm{P}\left(\mathcal{M}_k \mid x\right).$$

Posterior of Δ

The posterior of Δ is given by

$$f(\Delta \mid x) = \sum_k f(\Delta \mid \mathcal{M}_k, x) \, \mathrm{P}(\mathcal{M}_k \mid x),$$

where $f(\Delta \mid \mathcal{M}_k, x)$ is the posterior of Δ under model k. The posterior mean of Δ is

$$\begin{aligned} \mathrm{E}[\Delta \mid x] &= \int \Delta \sum_k f(\Delta \mid \mathcal{M}_k, x) \, \mathrm{P}(\mathcal{M}_k \mid x) \, d\Delta \\ &= \sum_k \mathrm{P}(\mathcal{M}_k \mid x) \underbrace{\int \Delta f(\Delta \mid \mathcal{M}_k, x) \, d\Delta}_{=\mathrm{E}[\Delta \mid x, \mathcal{M}_k]}, \end{aligned}$$

which is the BMA estimator of Δ.

Three Scenarios

The posterior probability

$$\mathrm{P}\left(\mathcal{M}_k \mid x\right) = \mathrm{P}\left(\mathcal{M}_k \text{ is the true model} \mid x\right)$$

can be strange if all candidate models are wrong, especially when we need to specify the prior probability that $\mathcal{M}_k$ is the true model.

1. The $\mathcal{M}$-closed setting means that one of the candidate models is the true data generating process.
2. The $\mathcal{M}$-complete setting means that but the true data generating process can be conceptualized, but it is not one of the candidate models due to, for example, model complexity or lack of information.
3. The $\mathcal{M}$-open setting means that the data generating process cannot be conceptualized and all candidate models are wrong.

Bayesian Stacking

Let $S(P,Q)$ be a scoring rule to measure the similarity between two probability measure P and Q. Let p and q be the corresponding densities. Then,

$$S(P,Q) \quad = \quad \int s(P,w)\,q(w)\,dw,$$

for some function $s(\cdot,\cdot)$. Bayesian stacking maximizes such similarity

$$S\left(\sum_{k} w_k f(\tilde{x} \mid x, \mathcal{M}_k),\ f_{\text{true}}(\tilde{x} \mid x)\right)$$

with respect to weights $\{w_k\}$ under the restriction that

$$\sum_{k} w_k = 1,\ 0 \leq 1 w_k \leq 1,\ \forall k.$$

Scoring Rule

Two commonly used scoring rules are

1. log score: $s(P,x)=\log p(x)$ such that $S(Q,Q)-S(P,Q)$ is the KL divergence.
   - Taking $p(x)=\sum_k w_k f(\tilde{x}\mid x,\mathcal{M}_k)$ is the same as maximizing the similarity between the stacked predictive distribution and the true predictive distribution.
2. energy score: $s(P,x)=\frac{1}{2}\mathrm{E}_P\left[\left\|X-\tilde{X}\right\|^{\beta}\right]-\mathrm{E}_P\left[\|X-x\|^{\beta}\right]$, where X and $\tilde{X}$ are two iid random variables, and the expectations are taken with respect to P.
   - If $\beta=2$, it reduces to $s(P,x)=-\left\|\mathrm{E}_P[X]-x\right\|^2$.
   - Maximizing the scoring rule is equivalent to minimizing the squared prediction error.

Leave-One-Out Cross Validation

However, we don't know $f_{\text{true}}\left(\tilde{x} \mid x\right)$ that is needed to evaluate

$$S\left(\sum_k w_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right),\ f_{\text{true}}\left(\tilde{x} \mid x\right)\right).$$

One alternative is to use leave-one-out cross validation as

$$\min_w \frac{1}{n}\sum_{i=1}^{n} s\left(\sum_k w_k f\left(x_i \mid x_{-i}, \mathcal{M}_k\right),\ x_i\right),$$

where

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int f\left(x_i \mid \theta_k, \mathcal{M}_k\right)\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k.$$

The stacked estimate of the predictive density is

$$\hat{f}\left(\tilde{x} \mid x\right) = \sum_k \hat{w}_k f\left(\tilde{x} \mid x, \mathcal{M}_k\right).$$

BMA and Bayesian Stacking

For BMA, it is alleged that, as $n \to \infty$,

- if one of the candidate models is the true model, say $\mathcal{M}_{k^*}$ is the true model,
- or, if all candidate models are misspecified and $\mathcal{M}_{k^*}$ has the smallest Kullback-Leibler divergence to the true model,

then $\mathrm{P}\left(\mathcal{M}_{k^*} \mid x\right) \to 1$.

In contrast to Bayesian model averaging,

- no prior $\mathrm{P}\left(\mathcal{M}_k\right)$ is needed in Bayesian stacking.
- Bayesian stacking is intended for the case where all candidate models are misspecified.

Importance Sampling

It is computationally intensive to compute the leave-one-out (LOO) predictive density

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right) d\theta_k$$

for each i, because we have to refit $\mathcal{M}_k$ n times to obtain all $\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)$.

Suppose that, for each k, we fit $\mathcal{M}_k$ using all the data and obtain L draws from the posterior $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$. Then,

$$f\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \int p\left(x_i \mid \theta_k, \mathcal{M}_k\right) \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \pi\left(\theta_k \mid x, \mathcal{M}_k\right) d\theta_k,$$

where the importance weight is

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)}.$$

Normalized Importance Sampling

We can rewrite the importance weight as

$$w_i\left(\theta_k, \mathcal{M}_k\right) = \frac{\pi\left(\theta_k \mid x_{-i}, \mathcal{M}_k\right)}{\pi\left(\theta_k \mid x, \mathcal{M}_k\right)} \propto \frac{f\left(x_{-i} \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid \mathcal{M}_k\right)}{f\left(x \mid \theta_k, \mathcal{M}_k\right) \pi\left(\theta_k \mid \mathcal{M}_k\right)} \propto \frac{1}{f\left(x_i \mid \theta_k, \mathcal{M}_k\right)}.$$

The normalized importance sampling estimator is

$$\hat{f}^{\mathrm{NIS}}\left(x_i \mid x_{-i}, \mathcal{M}_k\right) = \frac{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right) p\left(x_i \mid \theta_k, \mathcal{M}_k\right)}{\sum_{l=1}^{L} w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)},$$

where we sample $\theta_k^{(l)}$, $l = 1, \ldots, L$, from $\pi\left(\theta_k \mid x, \mathcal{M}_k\right)$.

However, the importance weights can be unstable if the distribution has a long tail that makes the importance weight very large.

Generalized Pareto Distribution

Theorem

Under suitable conditions on the random variable X, if the threshold u_0 is high enough, the conditional distribution of $X \mid X > u$ converges to a three-parameter generalized Pareto distribution (GPD), as $u \to \infty$. Its density is given by

$$f\left(x; u, \sigma, k\right) = \begin{cases} \frac{1}{\sigma}\left[1 + \frac{k(x-u)}{\sigma}\right]^{-1-1/k}, & \text{if } c \neq 0, \\ \frac{1}{\sigma}\exp\left(\frac{x-u}{\sigma}\right), & \text{if } c = 0, \end{cases}$$

for $y > u$ and $\sigma > 0$.

Pareto Smoothed Importance Sampling

Pareto Smoothed Importance Sampling stabilizes the large importance weights. Without loss of generality, suppose that $\left\{w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}$ has been ordered in increasing order.

- Consider the largest $N = \left\lfloor \min\left(0.2L,\, 3\sqrt{L}\right)\right\rfloor$ importance weights.
- We fit a GPD to $\left(w_i\left(\theta_k^{(L-N+1)}, \mathcal{M}_k\right), ..., w_i\left(\theta_k^{(L)}, \mathcal{M}_k\right)\right)$ with $u = w_i\left(\theta_k^{(L-N)}, \mathcal{M}_k\right)$.
- These N tail importance weights are replaced by

$$\min\left\{F^{-1}\left(\frac{z-1/2}{M}\right),\, \max_i w_i\left(\theta_k^{(l)}, \mathcal{M}_k\right)\right\}, \qquad z = 1, ..., M,$$

  where F^{-1} is the inverse distribution function of the fitted GPD. The other importance weights are unchanged.

Bayesian Statistics
Statistical Decision Theory

Shaobo Jin

Department of Mathematics

Basic Terminology

Let θ be an unknown quantity of interest. Θ is used to denote the set of all possible values of θ.

- If θ is a parameter in a statistical model, then Θ is the parameter space.

We will take a decision (or an action) d based on the observed data x, such as $d = \delta(x)$.

- The set $\mathcal{X}$ of all possible observations is called a sample space.
- The set $\mathcal{D}$ of all possible decisions is called a decision space.
- The function $\delta(x)$ is called a decision rule.

Decision Space: Example

Classification: Consider the problem of predicting $y_i \in \{0, 1\}$.

- The decision space is $\mathcal{D} = \{0, 1\}$ for 0-1 classification.
- The decision space is $\mathcal{D} = [0, 1]$ for probabilistic classification.

Estimation: Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be the parameter vector. We are interested in θ.

- The decision space is $\mathcal{D} = \Theta \subseteq \mathbb{R}^p$.

Prediction: Let $y \in \mathcal{X}$ be a future value that we want to predict.

- The decision space is $\mathcal{D} = \mathcal{X}$.

Loss and Risk

Definition (Loss function)

A loss function $L(\theta, d)$ is any non-negative function $L : \Theta \times \mathcal{D} \to [0, \infty)$.

For example:

$$\begin{aligned} L_2 \text{ loss}: &\qquad L(\theta - d) = (\theta - d)^2 \\ L_1 \text{ loss}: &\qquad L(\theta - d) = |\theta - d| \end{aligned}$$

Once we apply the loss function to the decision rule $\delta(x)$, we should treat $L(\theta, \delta(x))$ as a realization from the random variable $L(\theta, \delta(X))$.

Definition (Risk)

The (frequentist) risk is

$$R(\theta, \delta) = \mathrm{E}\left[L(\theta, \delta(X)) \mid \theta\right] = \int L(\theta, \delta(x)) f(x \mid \theta)\, dx.$$

Loss and Risk: Example

Example

Let $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ be a vector of iid random variables from the Bernoulli distribution $\text{Bernoulli}(p)$. We are interested in p.

- The sample space is $\mathcal{X} = [0, 1]$. The parameter space is $\Theta = [0, 1]$.
- The decision space is $\mathcal{D} = [0, 1]$.
- If we choose the loss function $L(\theta - d) = (\theta - d)^2$ and decision rule $\delta(X) = \bar{X}$, the risk is

$$R(\theta, \delta) = \mathrm{E}\left[L(p, \delta(X)) \mid p\right] \;\; = \;\; \mathrm{E}\left[\left(p - \bar{X}\right)^2 \mid p\right] = \frac{p(1-p)}{n},$$

where $\theta = p$ is treated as a fixed quantity here.

Integrated Risk

Definition (Integrated Risk)

The integrated risk is the expectation of the risk with respect to the prior $\pi(\theta)$:

$$\mathrm{E}\left[L(\theta,\delta)\right] = \int R(\theta,\delta)\,\pi(\theta)\,d\theta = \int \mathrm{E}\left[L(\theta,\delta(X)) \mid \theta\right]\pi(\theta)\,d\theta.$$

The decision that minimizes the integrated risk is called the Bayes decision rule (or Bayes estimator). The minimal integrated risk

$$\inf_{\delta} \mathrm{E}\left[L(\theta,\delta)\right]$$

is called the Bayes risk.

Find Bayes Decision

Let the posterior risk be

$$\mathrm{E}\left[L\left(\theta,\delta\right)\mid X=x\right] \;=\; \int L\left(\theta,\delta\right)\pi\left(\theta\mid x\right)d\theta.$$

Theorem (Find Bayes decision rule via posterior risk)

Suppose that

1. *there exists a decision rule with finite risk,*
2. *for almost all x, there exists a $\delta\left(x\right)$ minimizing the posterior risk $E\left[L\left(\theta,\delta\right)\mid X=x\right]$.*

Then, $\delta\left(x\right)$ is a Bayes decision rule.

Take-home Question: does the prior $\pi\left(\theta\right)$ have to be proper in order to apply this theorem?

Weighted L_2 Loss

Consider the weighted L_2 loss

$$L_W(\theta, d) = (\theta - d)^T W (\theta - d),$$

where W is a $p \times p$ symmetric and positive definite matrix.

Theorem

Suppose that there exists a decision rule with finite risk. Then, the Bayes decision rule with respect to the weighted L_2 loss is the posterior mean

$$\delta_B(X) = E[\theta \mid X = x],$$

where W does not depend on θ.

Find Bayes Decision: Example

Example

Consider the L_2 loss.

1. Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}(\theta)$. Suppose that $\theta \sim \text{Beta}(a, b)$. Find the Bayes decision rule.
2. Let X_1, ..., X_n be an iid sample from $N(\theta, 1)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. Find the Bayes decision rule.

Absolute Error Loss

For $k_1 > 0$ and $k_2 > 0$, define the absolute error loss

$$L_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d), & \text{if } \theta > d, \\ k_1(d - \theta), & \text{if } \theta \leq d. \end{cases}$$

If $k_1 = k_2$, such loss reduces to the L_1 loss.

Theorem

Suppose that there exists a decision rule with finite risk. Then, the Bayes decision rule δ_B with respect to the absolute error loss is the $k_2/(k_1 + k_2)$ fractile of the posterior distribution, i.e.,

$$P(\theta \leq \delta_B(x) \mid x) = \frac{k_2}{k_1 + k_2},$$

where k_1 and k_2 do not depend on θ. In particular, if $k_1 = k_2$, the Bayes rule is the posterior median.

Find Bayes Decision: Example

Example

Consider the L_1 loss.

1. Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}(\theta)$. Suppose that $\theta \sim \text{Beta}(a, b)$. Find the Bayes decision rule.
2. Let X_1, ..., X_n be an iid sample from $N(\theta, 1)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. Find the Bayes decision rule.

Prediction

Suppose that we want to predict a future observation, possibly from the conditional distribution $f(z \mid x, \theta)$. Let $L_{\text{pred}}(z, d)$ by the prediction error of predicting z by $d \in \mathcal{D}$.

- We can define the loss function as

$$L(\theta, d) = \int L_{\text{pred}}(z, d) f(z \mid x, \theta) \, dz.$$

- The integrated risk satisfies

$$\begin{aligned} \mathrm{E}\left[L_{\text{pred}}(z, \delta)\right] &= \int \int \int L_{\text{pred}}(z, \delta) f(z \mid x, \theta) \pi(\theta \mid x) m(x) \, dz dx d\theta \\ &= \int \left[\int \underbrace{\int L_{\text{pred}}(z, \delta) f(z \mid x, \theta) \, dz}_{=L(\theta,\delta)} \pi(\theta \mid x) \, d\theta \right] m(x) \, dx. \end{aligned}$$

Bayes Predictor

The Bayes predictor is the Bayesian decision rule that minimizes $\mathrm{E}\left[L_{\text{pred}}\left(z,\delta\right)\right]$.

- The posterior risk for prediction is

$$\begin{aligned}\int L\left(\theta,d\right)\pi\left(\theta\mid x\right)d\theta &= \int\left[\int L_{\text{pred}}\left(z,d\right)f\left(z\mid x,\theta\right)dz\right]\pi\left(\theta\mid x\right)d\theta\\ &= \int L_{\text{pred}}\left(z,d\right)f\left(z\mid x\right)dz,\end{aligned}$$

  where $f\left(z\mid x\right)$ is the density of the predictive distribution.
- Thus, $\delta\left(x\right)$ minimizing the posterior risk $\mathrm{E}\left[L_{\text{pred}}\left(z,\delta\right)\mid X=x\right]$ is the Bayes predictor.

L_2 Loss and L_1 Loss

Applying a previous theorem to the prediction case, we obtain the following Bayes predictors.

Theorem

Suppose that there exists a predictor with finite posterior risk.

1. *The Bayes predictor with respect to the weighted L_2 loss $L_{pred}(z, d) = (z-d)^T W (z-d)$ is the mean of the predictive distribution $E[Z \mid X = x]$, where W does not depend on θ.*
2. *The Bayes predictor with respect to the L_1 loss $L_{pred}(z, d) = |z-d|$ is the median of the predictive distribution.*

Find Bayes Predictor: Example

Example

Let Y_1, ..., Y_n be an iid sample from $N\left(\theta, 1\right)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. We want to predict an iid future observation $Z = Y_{n+1}$.

1. Find the predictive distribution.
2. Find the Bayes predictor under the L_2 loss.
3. Find the Bayes predictor under the L_1 loss.

$0-1$ Loss

Suppose that we are interested in a testing problem such that

$$\Theta = \Theta_0 \cup \Theta_1.$$

A nonrandomized test for a hypothesis is a statistic $\delta(X)$ taking values in $\{0,1\}$, where X is our data.

- $\delta = 1$ means that we reject H_0 and $\delta = 0$ means that we cannot reject H_0.
- Our decision space is $\mathcal{D} = \{0,1\}$.

We can define the $0-1$ loss by

$$L(\theta, d) = \begin{cases} 0, & \text{if } d = 0 \text{ and } \theta \in \Theta_0, \\ 0, & \text{if } d = 1 \text{ and } \theta \in \Theta_1, \\ 1, & \text{if } d = 0 \text{ and } \theta \in \Theta_1, \\ 1, & \text{if } d = 1 \text{ and } \theta \in \Theta_0, \end{cases} = \begin{cases} d, & \text{if } \theta \in \Theta_0, \\ 1-d, & \text{if } \theta \in \Theta_1. \end{cases}$$

Risk of 0-1 Loss

The frequentist risk is

$$\begin{aligned} R(\theta,\delta) &= \int L(\theta,\delta(x)) f(x \mid \theta)\, dx \\ &= \begin{cases} \mathrm{P}(\delta(X)=1), & \text{if } \theta \in \Theta_0, \text{ (just Type I Error probablity)} \\ \mathrm{P}(\delta(X)=0), & \text{if } \theta \in \Theta_1. \text{ (just Type II Error probablity)} \end{cases} \end{aligned}$$

The Bayes decision rule is

$$\delta(x) = \begin{cases} 1, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) < \mathrm{P}(\theta \in \Theta_1 \mid x), \\ 0, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) \geq \mathrm{P}(\theta \in \Theta_1 \mid x), \end{cases}$$

if $\mathrm{P}(\theta \in \Theta_0 \mid x) \in (0,1)$.

Loss for Distributions

Suppose that we want to find a distribution that fits the data well but we are less interested in the parameters themselves.

- Kullback-Leibler divergence (aka entropy loss):

$$L_{\mathrm{KL}} = \int \log\left(\frac{f\left(x \mid \theta\right)}{f\left(x \mid d\right)}\right) f\left(x \mid \theta\right) dx,$$

  where the truth is $f\left(x \mid \theta\right)$ and the decision is $f\left(x \mid d\right)$.

- Squared Hellinger distance:

$$\begin{aligned} L_{\mathrm{H}} &= \frac{1}{2}\int\left(\sqrt{\frac{f\left(x \mid d\right)}{f\left(x \mid \theta\right)}}-1\right)^2 f\left(x \mid \theta\right) dx \\ &= 1-\int\sqrt{f\left(x \mid d\right) f\left(x \mid \theta\right)}dx. \end{aligned}$$

Admissible Decision

Definition

A decision rule δ_0 is called inadmissible if there exits a decision rule δ_1 such that

$$\begin{aligned} R(\theta, \delta_0) &\geq R(\theta, \delta_1), \text{ for all } \theta \in \Theta, \\ R(\theta, \delta_0) &> R(\theta, \delta_1), \text{ for some } \theta \in \Theta. \end{aligned}$$

We say that δ_1 dominates δ_0. Otherwise, the decision rule δ_0 is called admissible.

- If $R(\theta, \delta_0) \geq R(\theta, \delta_1)$ for all θ, then the decision rule δ_0 is better than δ_1.
- If δ_0 is inadmissible, then δ_0 is uniformly dominated by another decision rule δ_1.

Admissible Decision: Example

Let X_1, ..., X_n be independent random variables where $X_i \sim N\left(\theta_i, 1\right)$. The parameter is $\theta = \begin{bmatrix} \theta_1 & \cdots & \theta_n \end{bmatrix}^T \in \mathbb{R}^n$.

- An unbiased estimator of θ is $\delta_0\left(X\right) = X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$.
- The James-Stein estimator is

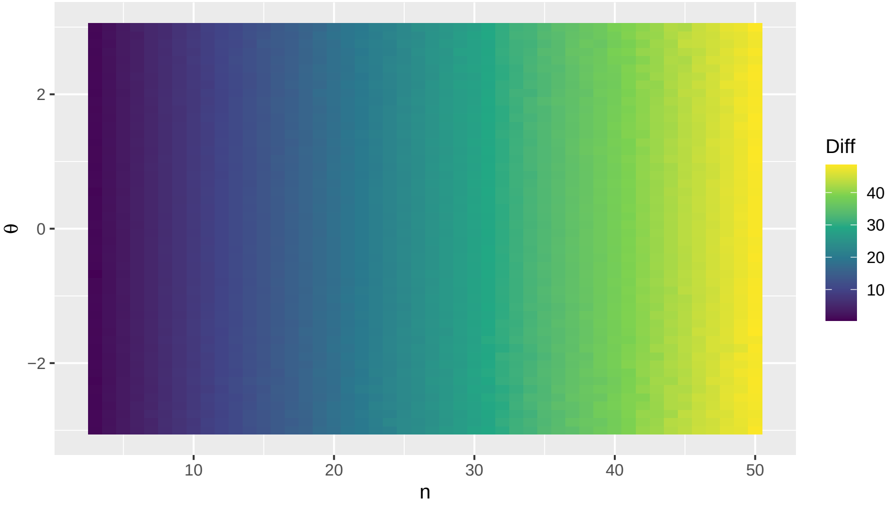
$$\delta_1\left(x\right) = \left(1 - \frac{n-2}{x^T x}\right) x.$$

If we consider the L_2 loss, then the difference in the risk satisfies

$$\mathrm{E}\left[L\left(\theta, \delta_0\left(X\right)\right) \mid \theta\right] - \mathrm{E}\left[L\left(\theta, \delta_1\left(X\right)\right) \mid \theta\right] \geq \frac{\left(n-2\right)^2}{n-2+\theta^T\theta} > 0,$$

for all θ.

Admissible Decision: Example



Minimax Decision Rule

Definition

A decision rule is minimax if it minimizes the maximum risk as

$$\inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} R\left(\theta, d\right) \right] = \inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} \mathrm{E}\left[L\left(\theta, d\left(X\right)\right) \mid \theta \right] \right].$$

Example

Suppose $X \mid \theta$ follows a 5-category multinomial distribution and $\theta \in \Theta = \{1, 2, 3\}$ indicates which distribution it is. The candidate distributions are

	x				
θ	1	2	3	4	5
1	0	0.05	0.05	0.8	0.1
2	0.05	0.05	0.8	0.1	0
3	0.9	0.05	0.05	0	0

Find Minimax Decision Rule: Example (Contd.)

Example

Suppose that our decision space $\mathcal{D} = \Theta$. Consider

Our decision rule					
	Observed x				
δ	1	2	3	4	5
δ_1	$d = 3$	3	2	2	1
δ_2	3	2	2	1	1
δ_3	1	1	1	1	1

Loss function			
	Decision d		
θ	1	2	3
1	$L(\theta, d) = 0$	0.8	1
2	0.3	0	0.8
3	0.3	0.1	0

Find the minimax decision rule.

Minimax and Admissible

Theorem (Relation between minimax rule and admissible rule)

1. *If there exists a unique minimax decision rule, then it is also admissible.*
2. *If δ is admissible and has constant risk, then δ is minimax.*
3. *Suppose that $\mathcal{D}$ is convex and, for all $\theta \in \Theta$, the loss function $L(\theta, \cdot)$ is strictly convex. If δ_0 is admissible and has constant risk, then δ_0 is unique minimax.*

Why Bayesian? 1: Bayes is Admissible

Theorem

The Bayes decision rule is admissible if either set of the following conditions hold.

1. $\pi(\theta) > 0$ *for all* $\theta \in \Theta$, $R(\theta, \delta)$ *is continuous in* θ *for all* δ, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta) \pi(\theta) d\theta \quad < \quad \infty.$$

2. *The Bayes decision rule is unique.*
3. $\mathcal{D}$ *is convex, the loss function* $L(\theta, \cdot)$ *is strictly convex for all* $\theta \in \Theta$, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta) \pi(\theta) d\theta \quad < \quad \infty.$$

Why Bayesian? 1: Bayes is Admissible

We can use the previous theorem to show an estimator is admissible.

Example

Let $X \sim N(\mu, 1)$ and the prior $\pi(\mu) = 1$. The parameter of interest is

$$\theta = 1(\mu \leq 0).$$

Consider a L_2 loss. Find the Bayes estimator of θ.

Blyth Theorem

Theorem

Let Θ be an open set. Suppose that the set of decision rules with continuous $R(\theta, d)$ in θ forms a class $\mathcal{C}$ such that for any $d' \notin \mathcal{C}$ we can find a $d \in \mathcal{C}$ such that d dominates d'. Let δ be an estimator such that $R(\theta, \delta)$ is continuous of θ. Let $\{\pi_n\}$ be a sequence of priors such that

1. $\int R(\theta, \delta) \pi_n(\theta) d\theta < \infty$ *for all* n,
2. *for every nonemptry open set* $\Theta_0 \in \Theta$, *there exist constants* $B > 0$ *and* N *such that*

$$\int_{\Theta_0} \pi_n(\theta) d\theta \quad \geq \quad B, \text{ for all } n \geq N,$$

3. $\int R(\theta, \delta) \pi_n(\theta) d\theta - \int R(\theta, \delta_n) \pi_n(\theta) d\theta \to 0$ *as* $n \to \infty$, *where* δ_n *is the Bayes rule under the prior* π_n.

Then, δ is admissible.

Limit of Bayes Rules

We have shown that the Bayes decision rule is admissible under some assumption. The Blyth theorem says that the admissible decision can be obtained such that

$$\lim_{n\to\infty} \int R\left(\theta,\delta\right)\pi_n\left(\theta\right)d\theta - \int R\left(\theta,\delta_n\right)\pi_n\left(\theta\right)d\theta \quad = \quad 0.$$

We can in fact claim that every admissible estimator is either a Bayes estimator or a limit of Bayes estimators as

$$\lim_{n\to\infty} \delta_n\left(x\right) \quad = \quad \delta_B\left(x\right), \text{ almost everywhere,}$$

under quite mild assumptions (e.g., $f\left(x \mid \theta\right) > 0$ for any $\left(x,\theta\right) \in \mathcal{X} \times \Theta$, $L\left(\theta,d\right)$ is continuous and strictly convex in d for every θ, among others).

Why Bayesian? 2: Bayes is Minimax

Definition

A prior distribution π is least favorable if

$$\int R(\theta, \delta) \pi(\theta) \, d\theta \quad \geq \quad \int R(\theta, \delta) \pi'(\theta) \, d\theta$$

for all prior distributions π'.

Theorem

Let δ_B be the Bayes decision rule with respect to the prior $\pi(\theta)$. Suppose that

$$\int R(\theta, \delta_B) \pi(\theta) \, d\theta \quad = \quad \sup_{\theta} R(\theta, \delta_B).$$

Then, δ_B is minimax and $\pi(\theta)$ is least favorable. Further, if δ_B is the unique Bayes decision rule with respect to the prior $\pi(\theta)$, then it is the unique minimax estimator.

Bayes is Minimax: A Corollary

Corollary

Let δ_B be the Bayes decision rule with respect to the proper prior $\pi(\theta)$. If δ_B has constant (frequentist) risk, then it is minimax.

Example

Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}(\theta)$. Suppose that $\theta \sim \text{Beta}(a, b)$. Find the minimax estimator of θ.

Bayes is Minimax: Another Corollary

Theorem

Suppose that δ_B is a Bayes decision rule with respect to a proper prior $\pi(\theta)$. If

$$R(\theta, \delta_B) \quad \leq \quad \int R(\theta, \delta_B)\, \pi(\theta)\, d\theta$$

for every $\theta \in \Theta$, then δ_B is minimax.

Minimax From Limit of Bayes Decision Rules

Theorem

Let $\{\pi_m\}$ be a sequence of proper prior distributions, and δ_m be the Bayes decision rule corresponding to the prior π_m. If δ is an estimator such that

$$\sup_{\theta} R(\theta, \delta) = \lim_{m \to \infty} \int R(\theta, \delta_m) \pi_m(\theta) \, d\theta.$$

Then δ is minimax.

Example

Let X_1, ..., X_n be iid observations from $N\left(\theta, \sigma^2\right)$, where σ^2 is known. Consider the L_2 loss $L(\theta, d) = (\theta - d)^2$. Find the minimax estimator.

Mutual Information

Let $m(x;\pi)$ be the marginal likelihood of x under the prior $\pi(\theta)$. We define the frequentist risk between $f(x \mid \theta)$ and $m(x;\pi)$ as

$$R_n(\theta,\pi) = \mathrm{KL}\left(f(x \mid \theta), m(x;\pi)\right) = \int f(x \mid \theta)\log\left[\frac{f(x \mid \theta)}{m(x;\pi)}\right]dx.$$

The integrated risk is then

$$\begin{aligned} R_n(\pi) &= \int R_n(\theta,\pi)\pi(\theta)\,d\theta = \int\int f(x,\theta)\log\left[\frac{f(x,\theta)}{m(x;\pi)\pi(\theta)}\right]dxd\theta \\ &= \mathrm{E}\left[\mathrm{KL}\left(\pi(\theta \mid x), \pi(\theta)\right)\right], \end{aligned}$$

which is the same as the mutual information of X and θ, and the expected Kullback-Leiber divergence.

Jeffreys Prior and Minimax

Suppose that some regularity conditions are satisfied, including Θ is a compact set, the Fisher information equals to the negative expected Hessian, among others.

- It has been proved that, among all positive and continuous priors,

$$\sup_{\pi} R_n\left(\pi\right) - \inf_{p(x)} \sup_{\theta\in\Theta} \mathrm{KL}\left(f\left(x \mid \theta\right), p\left(x\right)\right) \quad \to \quad 0.$$

- It has also been proved that the Jeffreys prior $\pi^*\left(\theta\right)$ is the unique continuous and positive prior such that

$$\sup_{\pi} R_n\left(\pi\right) - R_n\left(\pi^*\right) \quad \to \quad 0.$$

Hence, asymptotically, Jeffreys prior maximizes the mutual information, is the least favorable prior, and the integrated risk equals to the minimax risk.

Randomized Decision Rule

For simplicity, all results in our slides are formulated in terms of non-randomized decision rules. For completeness, we need to consider the randomized decision rules such that the action is generated according to some distribution once x has been observed.

Example

The Neyman-Pearson test is a randomized decision

$$\phi(x) = \begin{cases} 1, & \text{if } f_0(x) < kf_1(x), \\ r, & \text{if } f_0(x) = kf_1(x), \\ 0, & \text{if } f_0(x) > kf_1(x). \end{cases}$$

If $f_0(x) = kf_1(x)$, we let $\phi(x) = 1$ with probability r and $\phi(x) = 0$ with probability $1 - r$.

Loss and Risk For Randomized Decision

Since the decision is random, even though x is fixed, we need to take such extra randomness into account. That is, $\delta^*(x,\cdot)$ should be viewed as a density over $\mathcal{D}$ for fixed x.

1. The loss function of a randomized decision rule should be defined as an expected loss

$$L(\theta,\delta^*) = \int_{\mathcal{D}} L(\theta,a)\,\delta^*(x,a)\,da.$$

2. The (frequentist) risk is

$$\begin{aligned} R(\theta,\delta^*) &= \int L(\theta,\delta^*)\,f(x\mid\theta)\,dx \\ &= \int \left[\int_{\mathcal{D}} L(\theta,\delta^*(x,a))\,\delta^*(x,a)\,da\right] f(x\mid\theta)\,dx. \end{aligned}$$

Equivalence

The nonrandomized decision is a special case of the randomized decision rule, where we consider a dirac distribution $\delta^*(x, a) = 1$ on one action a. However, the inclusion of randomized decision rule does not affect the Bayes risk.

Theorem

For every prior π on Θ, the Bayes risk on the set of randomized decision rules is the same as the Bayes risk on the set of nonrandomized decision rules.

Bayesian Statistics
Asymptotic Theory

Shaobo Jin

Department of Mathematics

Bayesian Data Generating Process

In a Bayes model, the parameter θ is a random variable with known distribution π.

- Finding the true parameter makes no sense in a Bayes model.

Data that we observe are generated in a hierarchical manner:

$$\theta \sim \pi(\theta), \qquad X \mid \theta \sim f(x \mid \theta).$$

Given the data x, we can make inference for the data generating process.

In the usual frequentist statistics, we can find consistent estimators such that

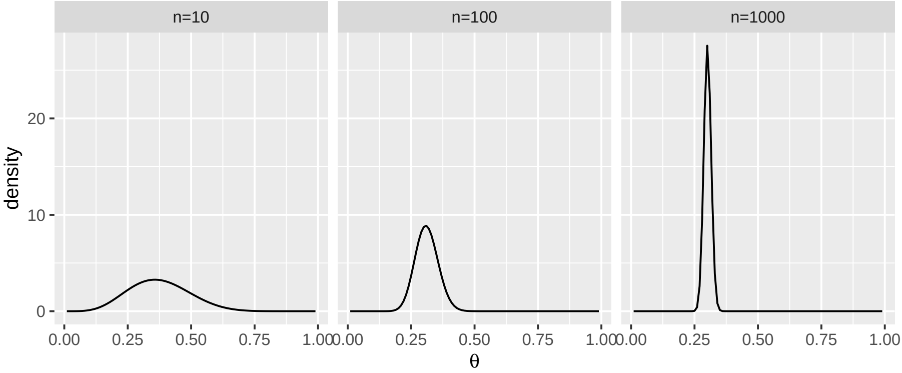
$$\mathrm{P}\left(\left\|\hat{\theta} - \theta\right\| < \epsilon\right) \quad\to\quad 1,$$

as $n \to \infty$. We also want the Bayes procedure to enable us to know θ with almost complete accuracy.

Example of Concentration

Example

Consider $X \mid \theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(a_0, b_0)$. The posterior is $\theta \mid x \sim \text{Beta}(a_0 + x, b_0 + n - x)$, which concentrates around the true θ_0 as $n \to \infty$.



Convergence in Probability

Let $X \in \mathbb{R}^p$ be a $p \times 1$ random vector of random variables.

Definition (Convergence in probability)

X_n converges in probability to X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}\left((X_n - X)^T (X_n - X) > \epsilon^2 \right) = 0.$$

It is denoted by $X_n \xrightarrow{P} X$. If X is a constant, then we also say X_n is consistent for X.

Example

Consider a sequence of independent random variables $\{X_n\}$, where $X_n \sim N\left(0,\ n^{-1}\right)$. Show that $X_n \xrightarrow{P} 0$.

Convergence Almost Surely

Definition (Convergence almost surely)

X_n converges almost surely to X if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1,$$

or equivalently, for every $\epsilon > 0$,

$$\mathrm{P}\left(\sqrt{\left(X_k - X\right)^T \left(X_k - X\right)} < \epsilon, \text{ for all } k \geq n\right) \to 1$$

It is denoted by $X_n \overset{a.s.}{\to} X$.

Example

Consider a sequence of independent random variables $\{X_n\}$, where $X_n \sim N\left(0,\ n^{-1}\right)$. Show that $X_n \overset{a.s.}{\to} 0$.

Some Useful Results for Us

Theorem

$X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{P} X$, but not the reverse.

Theorem (Slutsky Theorem)

1. *If $X_n \xrightarrow{P} X$ and $X_n - Y_n \xrightarrow{P} 0$, then $Y_n \xrightarrow{P} X$.*
2. *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $\begin{bmatrix} X_n \\ Y_n \end{bmatrix} \xrightarrow{P} \begin{bmatrix} X \\ Y \end{bmatrix}$.*

The theorem is also valid if every $\xrightarrow{P}$ is replaced by $\xrightarrow{a.s.}$.

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$. The theorem is also valid if $\xrightarrow{P}$ is replaced by $\xrightarrow{a.s.}$.

Law of Large Numbers

Theorem

Let X_1, X_2,... be iid random vectors, and let $\bar{X}_n = n^{-1}\sum_i X_i$. Then,

1. *Weak law of large numbers: If $E\left[\sqrt{X^T X}\right] < \infty$, then $\bar{X}_n \xrightarrow{P} \mu = E\left(X\right)$.*
2. *Strong law of large numbers: $\bar{X}_n \xrightarrow{a.s.} \mu$ for some μ if and only if $E\left[\sqrt{X^T X}\right] < \infty$ and $\mu = E\left(X\right)$.*

Consistency of MLE

Theorem

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$*. Suppose that the density is identified such that* $f(x \mid \theta) = f(x \mid \theta_0)$ *implies* $\theta = \theta_0$*. Assume*

C1 Θ *is an open set in* $\mathbb{R}^p$*, where* θ_0 *is an interior point,*

Then, under some other assumptions, the maximizer $\hat{\theta}$ *of* $\sum_{i=1}^{n} \log f(x_i \mid \theta)$ *is consistent, i.e.,* $\hat{\theta}_n \xrightarrow{P} \theta_0$*.*

If we change C1 to

C1' Θ *is a compact set in* $\mathbb{R}^p$*, where* θ_0 *is an interior point,*

then, under additional assumptions, $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$*.*

MAP Estimator

Suppose that data are generated from $X_i \mid \theta = \theta_0 \sim f\left(x \mid \theta_0\right)$, $i = 1, ..., n$. The MAP estimator essentially maximizes

$$\sum_{i=1}^{n} \log f\left(x_i \mid \theta\right) + \log \pi\left(\theta\right).$$

- Since the MLE of θ is a consistent estimator of the true value θ_0, the MAP should also be consistent if $n^{-1} \log \pi\left(\theta\right) \to 0$ as $n \to \infty$.
- We should also expect the MAP estimator to be strongly consistent, converging almost surely to θ_0.

Posterior Consistency

Apart from consistency of the estimator, we can also introduce consistency of the posterior distribution, as a frequentist evaluation of Bayesian posterior.

Definition

Suppose that data are generated from $X \mid \theta = \theta_0 \sim f(x \mid \theta_0)$.

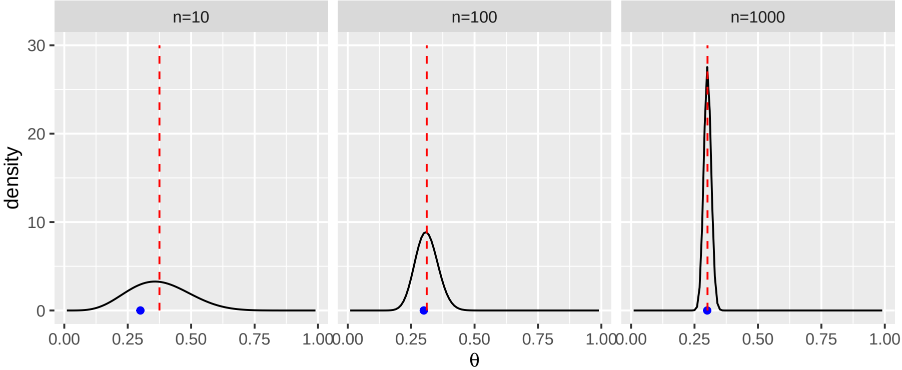
- the posterior is consistent at θ_0 if $\mathrm{P}(O \mid x)$ converges in probability to 1 under $f(x \mid \theta_0)$ as $n \to \infty$, for every open subset $O \subset \Theta$ with $\theta_0 \in O$.
- the posterior is strongly consistent at θ_0 if the convergence is almost surely. That is, for every open subset $O \subset \Theta$ with $\theta_0 \in O$, it holds that

$$\mathrm{P}(O \mid x) \to 1, \quad \text{as } n \to \infty, \quad \text{with probability } 1.$$

One Implication of Posterior Consistency

Posterior consistency suggests that, even though $\theta \sim \pi(\theta)$, the posterior $\pi(\theta \mid x)$ should concentrate around the θ_0 that generates the observed data.

- If $\pi(\theta \mid x)$ contracts to θ_0, we expect the Bayes estimator $\delta_B(x)$ should converge to θ_0.



Regularity Conditions

To establish such result, we need some regularity conditions. Let $L(\theta, d)$ be a loss function.

R1 There exists a constant $c_0 > 0$ for all d such that

$$c_0 \|d - \theta_0\| \quad \leq \quad L(\theta_0, d) - L(\theta_0, \theta_0).$$

- This condition implies that the loss function $L(\theta_0, \cdot)$ as a function of d has a minimum at $d = \theta_0$.

R2 There exists a constant K for all $X \sim f(x \mid \theta_0)$ such that

$$\int L^2(\theta, \theta_0) \pi(\theta \mid x) \, d\theta \quad \leq \quad K^2 \text{ almost sure.}$$

Consistency of Bayes Estimator

Theorem

Suppose that the loss function fulfills the conditions R1 and R2. Assume that

1. *for all $\epsilon > 0$ and all open sets $O \subset \Theta$ with $\theta_0 \in O$, it holds for*

$$B_\epsilon(\theta_0) = \{\theta : \theta \in O, |L(\theta, d) - L(\theta_0, d)| < \epsilon, \text{ for all } d\}$$

   that the prior probability $P(B_\epsilon(\theta_0)) > 0$ and there is an open subset in $B_\epsilon(\theta_0)$ such that θ_0 is an interior point.

2. *Let $X \sim f(x \mid \theta_0)$ and the sequence of posteriors $\pi(\theta \mid x)$ be strongly consistent at θ_0.*

Then, for $n \to \infty$, $\delta_B(x) \to \theta_0$ almost surely.

Consistency of General Estimator

Theorem

Suppose that the sequence of posteriors is strongly consistent at θ_0. Define the estimator $\hat{\theta}$ as the center of a ball of minimal radius that has posterior mass at least 0.5. Then $\hat{\theta}$ is consistent at θ_0.

Influence of Prior: Posterior

Theorem (Posterior robustness)

Consider $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$. *Let* θ_0 *be an interior point of* Θ, *and* π_1 *and* π_2 *be two prior densities, which are positive and continuous at* θ_0. *Let* $\pi_1(\theta \mid x)$ *and* $\pi_2(\theta \mid x)$ *be the respective posterior densities of* θ. *If* $\pi_1(\theta \mid x)$ *and* $\pi_2(\theta \mid x)$ *are both strongly consistent at* θ_0, *then*

$$\lim_{n\to\infty} \int \left|\pi_1(\theta \mid x) - \pi_2(\theta \mid x)\right| d\theta = 0, \text{ almost surely under } P_{\theta_0}.$$

Influence of Prior: Predictive Distribution

Theorem (Predictive robustness)

Assume that $\theta \mapsto P_\theta$ is one-to-one and continuous. Assume also that there is a compact set K such that $P(X \in K \mid \theta) = 1$ for all θ. Suppose that the posteriors $\pi_1(\theta \mid x)$ and $\pi_2(\theta \mid x)$ are both strongly consistent at θ_0, then the predictive distributions $\lambda_1(x^* \mid x)$ and $\lambda_2(x^* \mid x)$ satisfy

$$\lim_{n\to\infty} \left| \int \phi(x^*) \lambda_1(x^* \mid x) dx^* - \int \phi(x^*) \lambda_2(x^* \mid x) dx^* \right| = 0$$

for all bounded continuous functions ϕ.

Doob Consistency

Theorem (Doob's theorem for posterior consistency)

Suppose that $\theta \mapsto P(X \in A \mid \theta)$ *is one-to-one. Then, there exists a* $\Theta_0 \subseteq \Theta$ *with prior probability* $P(\Theta_0) = 1$ *such that, for every* $\theta_0 \in \Theta_0$, *if* $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$, *we have*

$$\lim_{n\to\infty} P(\theta \in O \mid X_1, \ldots, X_n) \;=\; 1, \; \textit{almost surely under } P_{\theta_0}$$

for any open set O *with* $\theta_0 \in O$.

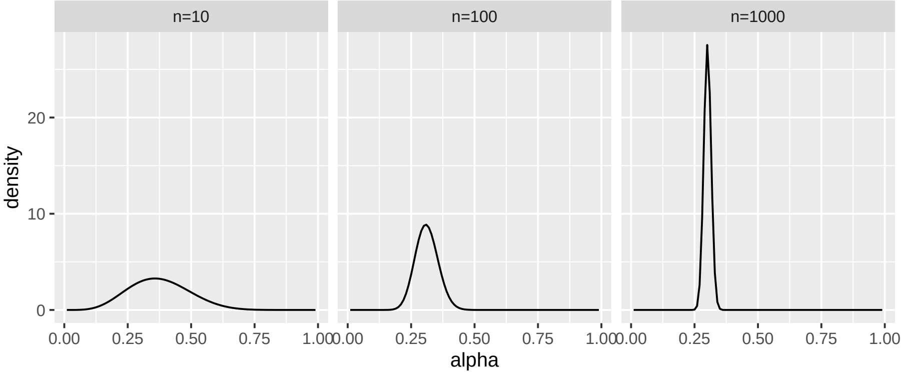
Doob's theorem says that the posterior will concentrate in a neighborhood, as long as

1. the statistics model is identified,
2. Θ_0 has strictly positive measure under the prior.

Example of Consistency

Example

Consider $X \mid \theta \sim \text{Binomial}\,(n, \theta)$ and $\theta \sim \text{Beta}\,(a_0, b_0)$. The posterior is $\theta \mid x \sim \text{Beta}\,(a_0 + x, b_0 + n - x)$. Show that posterior distributions concentrates around θ_0 as $n \to \infty$.



Positive Prior Assumption

The positive prior assumption plays a very important role in consistency. The posterior is

$$\pi(\theta \mid x) \quad \propto \quad f(x \mid \theta) \pi(\theta),$$

where $\pi(\theta \mid x) > 0$ only if $\pi(\theta) > 0$. We should never exclude any possible value from the prior.

Example

If $\pi(\theta) > 0$ for $\theta > 0$ and $\pi(\theta) = 0$ otherwise, then posterior can be better expressed as

$$\pi(\theta \mid x) \quad \propto \quad f(x \mid \theta) \pi(\theta) 1(\theta > 0).$$

The posterior is always zero for $\theta < 0$.

Doob's Theorem For Estimators

Theorem

Suppose that $\theta \mapsto P(X \in A \mid \theta)$ *is one-to-one. Then, there exists a* $\Theta_0 \subseteq \Theta$ *with prior probability* $P(\Theta_0) = 1$ *such that, for every* $\theta_0 \in \Theta_0$, *if* $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta_0)$, *we have*

$$\lim_{n\to\infty} E[g(\theta) \mid X_1, \ldots, X_n] = g(\theta_0), \text{ almost surely under } P_{\theta_0}$$

for any function $g(\theta)$ *such that*

$$\int g(\theta)\pi(\theta)\, d\theta < \infty.$$

For $g : \Theta \mapsto \mathbb{R}^p$, *the theorem holds to each component of* $g(\theta)$.

Doob's Theorem: Example

Example

Show that the posterior mean is strongly consistent.

1. Consider $X \mid \theta \sim \text{Binomial}\left(n, \theta\right)$ and $\theta \sim \text{Beta}\left(a_0, b_0\right)$. The posterior is $\theta \mid x \sim \text{Beta}\left(a_0 + x, b_0 + n - x\right)$.
2. Let X_1, ..., X_n be iid $N\left(\mu, \sigma^2\right)$, where σ^2 is known. Consider the prior $\mu \sim N\left(\mu_0,\ \sigma_0^2\right)$. The posterior is

$$\theta \mid x \quad \sim \quad N\left(\frac{\sigma_0^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2}, \quad \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}\right).$$

Limitations of Doob Consistency

Doob's theorem establishes consistency under quite mild conditions. However, it has been criticized based on various grounds.

1. It only guarantees consistency on set Θ_0 with prior probability $\mathrm{P}\left(\Theta_0\right) = 1$, not specific points θ_0.
2. It is less useful if θ is of infinite dimension, as the null set can be very large.

An alternative general theory is the Schwartz' theorem.

Distance/Divergence Between Two Distributions

Let P and Q be two probability measures with respective densities $p(x)$ and $q(x)$.

- Kullback-Leibler divergence (aka entropy loss):

$$\mathrm{KL}(\mathrm{P}, \mathrm{Q}) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\, dx.$$

- Hellinger distance:

$$\mathrm{H}^2(\mathrm{P}, \mathrm{Q}) = \frac{1}{2} \int \left[\sqrt{p(x)} - \sqrt{q(x)}\right]^2 dx = 1 - \mathrm{H}_{\frac{1}{2}}(\mathrm{P}, \mathrm{Q}),$$

where $\mathrm{H}_{\frac{1}{2}}(\mathrm{P}, \mathrm{Q}) = \int \sqrt{p(x)\, q(x)} dx$ is the Hellinger transform.

Schwartz' Theorem

Theorem

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$.

1. *KL condition: Suppose that $P(K_\epsilon(\theta_0)) > 0$ for all $\epsilon > 0$, where $K_\epsilon(\theta_0) = \{\theta : \ KL(P_{\theta_0}, P_\theta) < \epsilon\}$.*
2. *Hellinger condition: For every open set $O \in \Theta$ with $\theta_0 \in O$, there exist constants D_0 and $q_0 < 1$, such that*

$$H_{\frac{1}{2}}\left(P_{\theta_0}^{\otimes n}, P_{n,O^c}\right) \quad \leq \quad D_0 q_0^n,$$

   where P_{n,O^c} is defined by

$$P_{n,O^c}(A) \quad = \quad \int_A \int_{O^c} f_n(x \mid \theta) \frac{\pi(\theta)}{P(O^c)} d\theta dx.$$

Then, the sequence of posteriors is strongly consistent at θ_0.

Interpret the Conditions

1. The KL condition $\mathrm{P}\left(K_{\epsilon}\left(\theta_0\right)\right) > 0$ means that the prior does not exclude a neighborhood (in terms of the KL divergence) of θ_0.
2. The Hellinger condition means that the Hellinger distance

$$\mathrm{H}^2\left(\mathrm{P}_{\theta_0}^{\otimes n}, \mathrm{P}_{n,O^c}\right) = 1 - \mathrm{H}_{\frac{1}{2}}\left(\mathrm{P}_{\theta_0}^{\otimes n}, \mathrm{P}_{n,O^c}\right) \geq 1 - D_0 q_0^n.$$

Intuitively speaking, we can distinguish between $\mathrm{P}_{\theta_0}^{\otimes n}$ and $\mathrm{P}_{\theta}^{\otimes n}$ if θ is not in O where $\theta_0 \in O$.

- The Hellinger condition essentially replaces the identification condition ($\theta \mapsto \mathrm{P}\left(X \in A \mid \theta\right)$ is one-to-one) in Doob's theorem.

Hellinger Condition

Lemma

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$. Consider testing H_0: $P_{\theta_0}^{\otimes n}$ versus H_1: $\left\{P_\theta^{\otimes n} : \theta \in \Theta \setminus O\right\}$, where $O \in \Theta$ ise a neighborhood of θ_0. Suppose that there exists a nonrandomized test $\phi_n(x)$ and positive constants C and β such that

$$E[\phi_n(x) \mid \theta_0] + \sup_{\theta \in \Theta \setminus O} E[1 - \phi_n(x)] \quad \leq \quad C \exp(-n\beta).$$

Then the Hellinger condition holds.

Consistent Test

The condition in the lemma means that we can find a uniformly consistent test for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \in O^c$, where $O \in \Theta$ ise a neighborhood of θ_0.

- That is, there exists a test $\phi_n(x)$ such that

$$\mathrm{E}\left[\phi_n(x) \mid \theta_0\right] \to 0, \qquad \sup_{\theta \in O^c} \mathrm{E}\left[1 - \phi_n(x) \mid \theta\right] \to 0.$$

- If we can find a uniformly consistent test, we can apply the Hoeffding's inequality to obtain the exponential rate.
- The existence of a uniformly consistent test only requires that we can find a uniformly consistent estimator of θ, i.e.,

$$\sup_{\theta} \mathrm{P}\left[\left(\hat{\theta} - \theta\right)^T \left(\hat{\theta} - \theta\right) > \epsilon \mid \theta\right] \;\to\; 0.$$

Schwartz' Theorem: Another Version

Theorem

Let X_1, ..., X_n be iid from P_θ, denoted by $P_\theta^{\otimes n}$. Let $f_n(x \mid \theta)$ be the density of $x = (x_1, ..., x_n)$.

1. *KL condition: Suppose that $P(K_\epsilon(\theta_0)) > 0$ for all $\epsilon > 0$, where $K_\epsilon(\theta_0) = \{\theta : \ KL(P_{\theta_0}, P_\theta) < \epsilon\}$.*
2. *Uniformly consistent test condition: Let $O \in \Theta$ be a neighborhood of θ_0. Consider testing H_0: $\theta = \theta_0$ versus H_1: $\theta \in O^c$. There exists a test $\phi_n(x)$ such that*

$$E[\phi_n(x) \mid \theta_0] \to 0, \qquad \sup_{\theta \in O^c} E[1 - \phi_n(x) \mid \theta] \to 0.$$

Then, the sequence of posteriors is strongly consistent at θ_0.

Consistency and Normality of MLE

Theorem

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x \mid \theta)$. *Assume that*

- C1 Θ *is an open set in* $\mathbb{R}^p$, *where* θ_0 *is an interior point,*
- C2 $\{x : f(x \mid \theta) > 0\}$ *does not depend on* θ, *i.e., common support,*
- C3 $\int f(x \mid \theta)\, dx$ *can be twice differentiable under the integral sign,*
- C4 *The Fisher information* $\mathcal{I}(\theta)$ *satisfies* $0 < I(\theta) < \infty$.

If some other regularity conditions are satisfied, then there exists a strongly consistent sequence $\hat{\theta}$ *of roots of the likelihood equation*

$$\frac{\partial \sum_{i=1}^{n} \log f(x_i \mid \theta)}{\partial \theta} = 0,$$

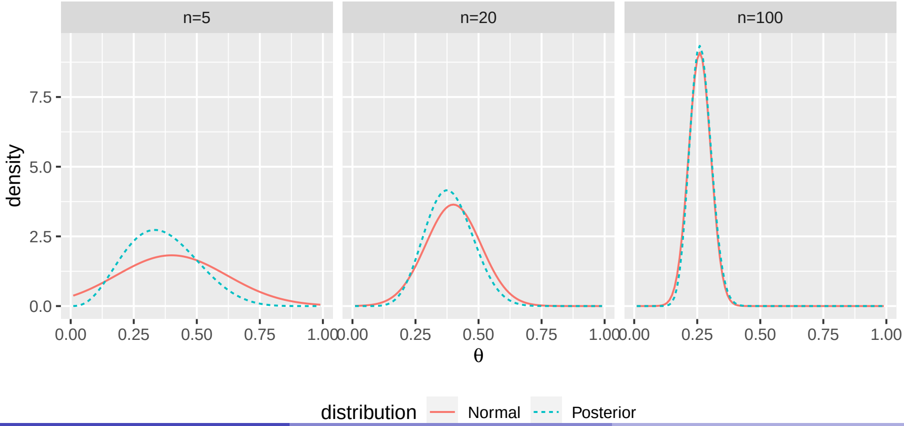
such that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1}(\theta)\right).$$

Posterior Distribution

The posterior of a beta-binomial model is

$$\text{Beta}\left(a_0 + \sum_{i=1}^{n} x_i, b_0 + n - \sum_{i=1}^{n} x_i\right).$$



Normality of Posterior

The heuristic argument that we aim to conclude is that posterior distributions in differentiable parametric models converge to the Gaussian posterior distribution.

- If $\hat{\theta}$ is the MLE of θ, then

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \ \overset{d}{\to} \ N\left(0, \mathcal{I}^{-1}\left(\theta\right)\right).$$

- We want to claim that the difference between the posterior distribution $\pi\left(\theta \mid x_1, ..., x_n\right)$ and the normal distribution

$$\hat{\theta} \ \approx \ N\left(\theta, \frac{1}{n}\mathcal{I}^{-1}\left(\theta\right)\right)$$

converge to zero.

Bernstein-von Mises Theorem

Let $\hat{\theta}$ be the strongly consistent sequence of roots of the likelihood equation. Define $t = \sqrt{n}\left(\theta - \hat{\theta}\right)$. Let $\pi\left(t \mid x_1, ..., x_n\right)$ be the posterior density of t.

Theorem

Let $X_1, \ ..., \ X_n \overset{iid}{\sim} f\left(x \mid \theta\right)$. Suppose that the assumptions C1 - C4 in the previous theorem hold. Assume that $\pi\left(\theta\right)$ is continuous and $\pi\left(\theta\right) > 0$ for all $\theta \in \Theta$.

1. *If some other regularity conditions are satisfied, then*

$$\left|\pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)\right| \overset{a.s.}{\to} 0 \text{ under } P_\theta,$$

   where $\phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)$ is the density of $N\left(0, \mathcal{I}^{-1}\left(\theta\right)\right)$.

2. *If, in addition, $\mathcal{I}\left(\theta\right)$ is continuous, then,*

$$\left|\pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\hat{\theta}\right)\right)\right| \overset{a.s.}{\to} 0 \text{ under } P_\theta.$$

Bernstein-Von Mises Theorem: Example

Example

Suppose that $X_1,\ldots, X_n$ are iid Bernoulli$\left(\theta\right)$. We consider a continuous prior $\pi\left(\theta\right) > 0$ for all $\theta \in \Omega$. Approximate the posterior of θ.

Total Variation Distance

Let P and Q be two probability measures. Then, their total variation distance is

$$\sup_{A} \left|\mathrm{P}\left(A\right) - \mathrm{Q}\left(A\right)\right|,$$

for all Borel sets A. If p and q are the respective densities, then,

$$\sup_{A} \left|\mathrm{P}\left(A\right) - \mathrm{Q}\left(A\right)\right| = \frac{1}{2}\int \left|p\left(x\right) - q\left(x\right)\right| dx.$$

The Bernstein-von Mises theorem indicates that

$$\begin{aligned} \sup_{A} \left|\mathrm{P}\left(t \in A \mid x_1, ..., x_n\right) - \mathrm{P}\left(t \in A \mid t \sim \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)\right)\right| &\overset{a.s.}{\to} 0, \\ \text{and } \int \left|\pi\left(t \mid x_1, ..., x_n\right) - \phi\left(t, 0, \mathcal{I}^{-1}\left(\theta\right)\right)\right| dt &\overset{a.s.}{\to} 0. \end{aligned}$$

Bayesian Credible Set

For simplicity, the classic one dimensional MLE $\hat{\theta}$ satisfies that $\mathrm{P}\left(\theta \in C\left(\alpha\right)\right) \to 1-\alpha$,where

$$C\left(\alpha\right) = \left[\hat{\theta} - \lambda_{1-\alpha/2}\sqrt{\frac{\mathcal{I}^{-1}\left(\theta\right)}{n}}, \hat{\theta} + \lambda_{1-\alpha/2}\sqrt{\frac{\mathcal{I}^{-1}\left(\theta\right)}{n}}\right].$$

The Bernstein-von Mises theorem allows us to approximate the posterior probability. In particular, let

$$B\left(\alpha\right) = \left\{\theta : \pi\left(\theta \mid x\right) \geq c_n\right\}$$

such that $\mathrm{P}\left(B\left(\alpha\right) \mid x\right) = 1-\alpha$. Then, for any $\epsilon > 0$,

$$\mathrm{P}\left(C\left(\alpha+\epsilon\right) \subset B\left(\alpha\right) \subset C\left(\alpha-\epsilon\right)\right) \to 1.$$

Example

Approximate the Bayesian credible set in the beta-binomial model.

Asymptotic Efficiency of Bayes Estimators

Consider the squared loss. The Bayes estimator of θ is the posterior mean $\tilde{\theta} = \mathrm{E}\left[\theta \mid x\right]$. The Bernstein-von Mises theorem may also indicate that

$$\mathrm{E}\left[\sqrt{n}\left(\theta - \hat{\theta}\right) \mid x\right] \;\; \rightarrow \;\; 0.$$

This suggests that

$$\sqrt{n}\left(\tilde{\theta} - \hat{\theta}\right) \;\; \rightarrow \;\; 0.$$

- The Bayes estimator and the MLE are asymptotically equivalent.
- The Bayes estimator is asymptotically efficient since MLE is asymptotically efficient.

An Counterexample

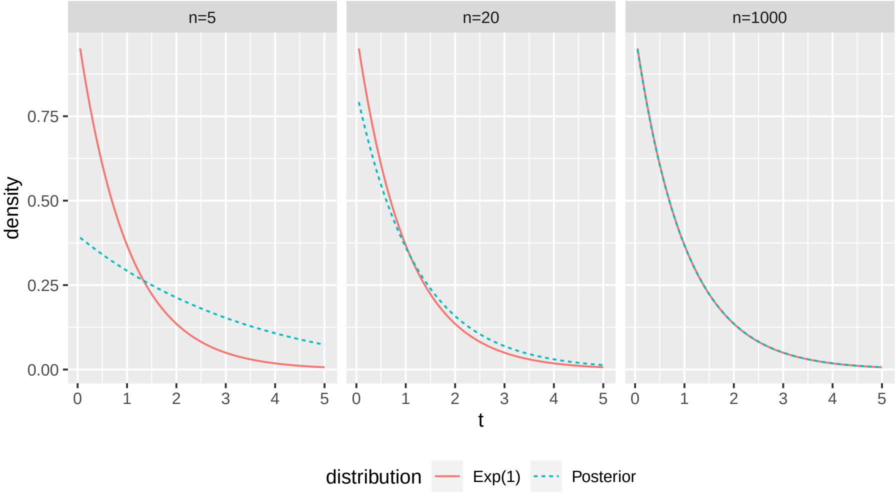
Example

Suppose that X_1, ..., X_n are iid with density

$$f\left(x \mid \theta\right) = \exp\left\{-\left(x-\theta\right)\right\}, \quad x > \theta.$$

The prior is $\theta \sim \text{Gamma}\left(2, b_0\right)$. Find the posterior of θ and show that the Bernstein-von Mises theorem is not applicable.

Counterexample: Posterior of Shifted Exponential


n=5
n=20
n=1000
density
t
distribution
Exp(1)
Posterior