Elementary Probability Theory
1. Conditional Probability
2. Bayesian Theorem
Transformation of Variable
1. Injection Function (Jacobian Method)
2. Non-Injection Function (CDF->PDF Method)
Common Probability Theory Tool
1. Double Expectation Formula(Law of Total Expectation)
2. Chain Rule of Probability Theory
Prior and Conjugate Prior
1. Conjugate Prior of Elementary Distribution
2. The Jeffery's Prior
The Bayesian Inference
1. Point Estimation
2. Interval Estimation
3. The Hypothesis Testing
The Multi-parameter Model
Bayesian Computation
1. Sampling
2. Markov Chain Monte Carlo(MCMC)
The Asymptotic Theory
The Hierarchical Model
Exchangeability

Elementary Probability Theory

1. Conditional Probability

Given events A and B, we have the **Conditional Probability Formula**:

$$P(A|B) = \frac{P(AB)}{P(B)} \text{ if P(B)} \neq 0$$

Exercise

An urn contains 10 red balls and 15 white balls. You pick two balls at random without replacement.

1. What is the probability that first ball is red?
2. What is the probability that second ball is red?
3. What is the probability that both ball are white?
4. What is the probability that the second ball is red given that first ball is white?
5. What is the probability that the first ball is red given that second ball is white?

Solution

Let A be first ball is red and let F be second ball is red.

1. $P(E) = \frac{10}{25} = 0.4$
2. $P(F) = P(F|E)P(E) + P(F|\bar{E})P(\bar{E}) = \left(\frac{9}{24}\right)\left(\frac{10}{25}\right) + \left(\frac{10}{24}\right)\left(\frac{15}{25}\right) = \frac{240}{600} = \frac{2}{5}$
3. $P(\bar{E} \cap \bar{F}) = P(\bar{F}|\bar{E})P(\bar{E}) = \left(\frac{14}{24}\right)\left(\frac{15}{25}\right) = \frac{210}{600} = \frac{7}{20}$
4. $P(F|\bar{E}) = \frac{10}{24} = \frac{5}{12}$
5. $P(E|\bar{F}) = \frac{P(E \cap \bar{F})}{P(\bar{F})} = \frac{P(\bar{F}|E)P(E)}{P(\bar{F})} = \frac{\left(\frac{15}{24}\right)\left(\frac{2}{5}\right)}{1-\frac{2}{5}} = \frac{\left(\frac{15}{24}\right)\left(\frac{2}{5}\right)}{\frac{3}{5}} = \frac{\frac{30}{120}}{\frac{3}{5}} = \frac{5}{12}$

2. Bayesian Theorem

Given events A, B, we have the Bayesian Theorem:

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

For continuous case, we have:

$$p(x|y) = \frac{\pi(x)p(y|x)}{p(y)}$$

Exercise

In the world of Three Body there is a kind of era called chaos era, during which the sun will not rise in most of the days in the era. Let's assume that the sun only rises 5 days out of the era (149 days). An experienced player of the Three Body game says: "the sun will rise tomorrow!". By experience, he can correctly predict 95% of the time when the sun does rise and correctly predicts no rising 90% of the time when the sun does not rise. What is the probability the sun will actually rise tomorrow?

Solution

$P(Rise) = \frac{5}{149}$

$P(PredictedRise|Rise) = 0.95$

$P(PredictedNoRise|NoRise) = 0.9$

$P(Rise|PrediectedRise) = \frac{P(Rise)P(PredictedRise|Rise)}{P(PredictedRise)} = \frac{P(Rise)P(PredictedRise|Rise)}{P(PredictedRise|Rise)P(Rise)+P(PredictedRise|NoRise)P(NoRise)} = 0.8642$

Transformation of Variable

1. Injection Function (Jacobian Method)

The basic rule for transformation of densities considers an invertible, smooth mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ with inverse $f^{-1} = g$, i.e. the composition $g \circ f(\mathbf{z}) = \mathbf{z}$. If we use this mapping to transform a random variable z with distribution $q(\mathbf{z})$, the resulting random variable $\mathbf{z}' = f(\mathbf{z})$ has a distribution :

$$q\left(\mathbf{z}'\right) = q(\mathbf{z})\left|\det \frac{\partial f^{-1}}{\partial \mathbf{z}'}\right| = q(\mathbf{z})\left|\det \frac{\partial f}{\partial \mathbf{z}}\right|^{-1}$$

where the last equality can be seen by applying the chain rule (inverse function theorem) and is a property of Jacobians of invertible functions.

Exercise

A prior is considered flat if it is proportional to a constant.

$$p(\theta) \propto c$$

Show that if we transform the parameter such that the new parameter $\mathit{\Theta} = exp(\theta)$, a Beta(1,1) prior on θ is no longer flat on $\mathit{\Theta}$

Solution

$$p(\Theta) = p(\theta)|det\frac{\partial e^{\theta}}{\partial \theta}|^{-1} = p(\theta) * \frac{1}{e^{\theta}} = \frac{1}{\Theta}$$

Which is not proportional to constant, so it is not flat on Θ

2. Non-Injection Function (CDF->PDF Method)

This time when we face a non injection map like $f(X) = X^2$, we should use CDF to PDF Method

Exercise

We'll use the inverse transform algorithm, $X = F^{-1}(U)$. To find X, solve the equation m $F(X) = U$ for X,

$$1 - \left(\frac{5}{X}\right)^2 = U \quad \Leftrightarrow \quad \left(\frac{5}{X}\right)^2 = 1 - U \quad \Leftrightarrow \quad \frac{5}{X} = \sqrt{1-U} \quad \Leftrightarrow \quad X = \frac{5}{\sqrt{1-U}}.$$
$$U = 0.36,\ X = \frac{5}{\sqrt{1-0.36}} = \frac{5}{0.8} = 6.25.$$

Since $(1-U)$ is also a standard uniform random variable, the answer $X = \frac{5}{\sqrt{U}} = \frac{5}{\sqrt{0.36}} = 8.33$ is just as good. It's useful to check that both values of X belong to the range of possible values of this Pareto distribution, $X \geq 5$.

Common Probability Theory Tool

1. Double Expectation Formula(Law of Total Expectation)

Given u,v are random variable vector, we have:

$$E[u] = E[E[u|v]]$$

Exercise

Given a random process $\{Z_n; n >= 1\}$, which is a **martingale**,i.e $Z_{k-1} = E[Z_k|Z_{k-1}, Z_{k-2}, \cdots, Z_1]$

Show that $E[Z_2|Z_0] \geq Z_1$, assume $E[Z_1|Z_0] = Z_0$

Solution

With Law of Total Expectation $E[Z_2|Z_1] = E[E[Z_2|Z_1]|Z_0] = E[Z_1|Z_0] = Z_0$

2. Chain Rule of Probability Theory

Given u,v,w, we have:

$$f(u, v, w) = f(u|v, w)f(v, w) = f(u|v, w)f(v|w)f(w)$$

Prior and Conjugate Prior

1. Conjugate Prior of Elementary Distribution

Let F represent the distribution family of the prior $\pi(\theta)$, if $\forall x, \pi \in F, f(\theta|x) \in F$, then we call F is a **conjugate prior** of the likelihood distribution.

- Beta Distribution as a conjugate Prior

  Eg. The beta distribution Beta(a,b) is the conjugate prior of binomial distribution Bino(n,θ)

  Prove:

  $$p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} = \theta^{a+x-1}(1-\theta)^{b+n-x-1} \sim Beta(a+x, b+n-x)$$

- Gamma Distribution as a conjugate Prior

  Eg. The gamma distribution Gamma(a,b) is the conjugate prior of poisson distribution Poisson(θ)

  Prove:

  $$p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta^{a-1}e^{-b\theta}\theta^x e^{-\theta} = \theta^{a+x-1}e^{-(b+1)\theta}$$

  The similar case is when $X_1, \cdots, X_n \sim Poisson(\theta)$

- Normal Distribution as a conjugate Prior

  Theorem: If $X_1, \cdots, X_n \sim N(\theta, \tau^2), \theta \sim N(\mu, \sigma^2)$, (variance is known) then the Posterior Distribution $f(\theta|x) \sim N(\hat{\mu}, \hat{\tau}^2)$

  in which

  $$\hat{\mu} = \frac{u\tau^2/n + \bar{X}\sigma^2}{\sigma^2 + \tau^2/n}, \hat{\tau}^2 = \frac{\tau^2/n * \sigma^2}{\tau^2/n + \sigma^2}$$

2. The Jeffery's Prior

- Definition

  Jeffreys principle leads to defining the noninformative prior density as $p(\theta) \sim [J(\theta)]^{1/2}$, where $J(\theta)$ is the Fisher information for θ:

$$J(\theta) = E\left(\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2|\theta\right) = -E\left(\frac{d^2\log p(y|\theta)}{d\theta^2}|\theta\right)$$

- Calculation
  - Log Likelihood
  - Fisher Information Matrix
  - Find Jeffery's Prior

Exercise

Find the Jeffery's Prior for the Binomial Distribution Bino(n,p)

Solution

See lecture 5 Example 1. See last RC for multi case. (I guess multi case will not be covered in exam)

The Bayesian Inference

1. Point Estimation

- Posterior Mean
  - Use conjugate prior to find posterior distribution and get it
- Posterior Median
  - Most difficult one, you should use integration to get it. **Pay Attention to Symmetric Distribution**
- Posterior Modal Estimation(MAP)
  - Use method like MLE (take derivative on log of posterior distribution) to get it

Exercise

Suppose we have observed X1,X2,⋯ ,Xn from exp(λ) distribution and λ ~ Gamma(α,β).

Derive the posterior mean estimator of λ as a function of α,β, what happens when n gets larger?

Solution

$$p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta^{a-1}e^{-b\theta}\prod_{i=1}^{n}\theta e^{-\theta x_i} = \theta^{a+n-1}e^{-(b+\sum_i x_i)\theta} \sim Gamma(a+n, b+\sum_i x_i)$$

The posterior mean estimator is $\frac{a+n}{a+n+b+\sum_i x_i}$.

For MAP estimator you can do exercise on your own, very easy.

2. Interval Estimation

- Credible Interval

  It is just the definition you miss understood in VE401. There is xx possibility that θ fall in the interval.
- HPD Interval

  Highest possibility interval. Pay attention to the symmetric or monotonous property of PDF

3. The Hypothesis Testing

The task is more straightforward. Compute $\alpha_0 = P(\Theta_0|x)$ and $\alpha_1 = P(\Theta_1|x)$ **They are normalized values!!!!**

if $\alpha_0 > \alpha_1$: Accept H_0

if $\alpha_1 > \alpha_0$: Accept H_1

if α_0 is close to α_1: Hard to say, better adjust prior or collect more data.

The ratio of α_0/α_1 is called the posterior odds ratio of H_0 to H_1, and π_0/π_1 is called the prior odds ratio. The quantity

$$B = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

is called the Bayes factor if favor of Θ_0

Different Kinds Of Cases

- H0:$\theta = \theta_0$,H1:$\theta = \theta_1$

  We only consider discrete random variables.

$$\alpha_0(x) = \frac{\rho(x|\theta_0)\pi(\theta_0)}{m(x)}$$

$$\alpha_1(x) = \frac{\rho(x|\theta_1)\pi(\theta_1)}{m(x)}$$

  $\frac{\alpha_0(x)}{\alpha_1(x)} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)}$, we compare this ratio with 1.

$$B(x) = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

- H0 : θ $\in \mathit{\Theta}_0$ versus H1 : θ $\in \mathit{\Theta}_1$
  - Discrete random variable

$$\alpha_0(x) = \sum_{\theta_i \in \Theta_0} f(\theta_i|x)$$

$$\alpha_1(x) = \sum_{\theta_i \in \Theta_1} f(\theta_i|x)$$

  $\frac{\alpha_0(x)}{\alpha_1(x)} > 1$: accept H_0, otherwise, accept H_1

$$B(x) = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}$$

  - Continuous random variable

$$\alpha_0(x) = \int_{\theta \in \Theta_0} f(\theta|x) d\theta$$

$$\alpha_1(x) = \int_{\theta \in \Theta_1} f(\theta|x) d\theta$$

    $\frac{\alpha_0(x)}{\alpha_1(x)} > 1$: accept H_0, otherwise, accept H_1

$$B(x) = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1},\ \pi_0 = \int_{\Theta_0} \pi(\theta) d\theta,\ \pi_1 = \int_{\Theta_1} \pi(\theta) d\theta$$

- H0 : θ = θ_0 versus H1 : θ $\neq \theta_1$
  - θ is discrete random variable (Similar to previous derivations)
  - θ is continuous random variable
    To prevent zero probability $H_0 : \theta \in [\theta_0 - \epsilon, \theta_0 + \epsilon]$ versus $H_1 : \theta \neq \theta_1$
    Or Just assign $\pi(\theta_0) = \pi_0$

$$\pi(\theta) = \pi_0 l_{\theta_0}(\theta) + \pi_1 g_1(\theta)$$

  g_1 is a discontinuous probability density function. For marginal distribution:

$$p(x) = \int p(x|\theta)\pi(\theta) d = \pi_0 p(x|\theta_0) + \pi_1 p_1(x)$$

$$p_1(x) = \int_{\theta \neq \theta_0} p(x|\theta) g_1(\theta)$$

-

Exercise

The annual number of forest fires in a certain county in California has Poisson distribution with parameter θ, independently of other years. During three consecutive years, there were 0, 1 , and 0 forest fires. Assume an improper non-informative prior$\pi(\theta) = \frac{1}{\theta}$

Is there a significant evidence that θ, the annual frequency of forest fires, does not exceed 1 fire per year?

Solution

SOLUTION. We need to test $H_0 : \theta > 1$ vs $H_A : \theta \le 1$ (because a significant evidence is only needed to reject H_0 in favor of H_A.

For $\boldsymbol{X} = (X_1, X_2, X_3) = (0, 1, 0) \sim$Poisson$(\theta)$,

$$f(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{3} \frac{e^{-\theta}\theta^{X_i}}{X_i!} \sim e^{-\theta}\theta^0 \cdot e^{-\theta}\theta^1 \cdot e^{-\theta}\theta^0 = e^{-3\theta}\theta.$$

Then, the posterior density is

$$\pi(\theta \mid \boldsymbol{X}) \sim f(\boldsymbol{X} \mid \theta)\pi(\theta) \sim e^{-3\theta}\theta \cdot \frac{1}{\theta} = e^{-3\theta}.$$

(3' 算出 Posterior)

This is **Exponential** density with parameter $\lambda = 3$ and cumulative distribution function $F(\theta \mid \boldsymbol{X}) = 1 - e^{-3\theta}$. Then,

$$\begin{aligned} \boldsymbol{P}\{H_A \mid \boldsymbol{X}\} &= \boldsymbol{P}\{\theta \le 1\} = F(1 \mid \boldsymbol{X}) = 1 - e^{-3} = 0.9502 \\ \boldsymbol{P}\{H_0 \mid \boldsymbol{X}\} &= \boldsymbol{P}\{\theta > 1\} = 1 - F(1 \mid \boldsymbol{X}) = e^{-3} = 0.0498 \end{aligned}$$

(2' α_1)

(2' α_0)

With equal losses, we have a sufficient evidence supporting H_A. So, yes, there a significant evidence that the annual frequency of forest fires does not exceed 1 fire per year.

结论 1'

The Multi-parameter Model

- Joint Posterior Density

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)\rho(\theta_1, \theta_2).$$

- Marginal density

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2.$$

Example. $y \sim N(\mu, \sigma^2)$ and non-informative prior $\pi(\mu, \sigma^2) \propto \sigma^{-2}$.

$$\begin{aligned} \rho\left(\mu, \sigma^2|y\right) &\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \\ &= \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right) \\ &= \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y} - \mu)^2\right]\right). \end{aligned}$$

Example. $y \sim N(\mu, \sigma^2)$ ànd non-informative prior $\pi(\mu, \sigma^2) \propto \sigma^{-2}$.

$$p(\sigma^2|y) = \int p(\mu, \sigma^2|y)d\mu \propto \left(\sigma^2\right)^{-(n+1)/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right).$$

Known Mean: $\sigma^2|y \sim \text{Inv}{-}\chi^2(n, v), v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$.

Unknown Mean: $\sigma^2|y \sim \text{Inv} - \chi^2(n-1, s^2), s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.
Posterior predictive distribution:

$$p(\tilde{y} \mid y) = t_{n-1}\left(\tilde{y} \mid \bar{y}, \left(1 + \frac{1}{n}\right)s^2\right).$$

Bayesian Computation

1. Sampling

- Gird Sampling

  Grid Sampling Steps.

  1. Create an even-spaced grid: $g_1 = a + i/2, \cdots, g_m = b - i/2$ where a is the lower,and b is the upper limit of the interval on which we want to evaluate the posterior, i is the increment of the grid, and m is the number of grid numbers.
  2. Evaluate values of the unnormalized posterior density in grid points $q(g_1; y), \cdots, q(g_m; y)$ and normalize them to obtain estimated values of the $q(g_1; y), \cdots, q(g_m; y)$ and hormalze the: $\hat{p}_i = \frac{q(g_i;y)}{\sum_{i=1}^{m} q(g_i;y)}.array$
  3. For every $s = 1, \cdots, S$, generate λ_s from a categorical distribution with outcomes

  $g_1, \cdots, g_m$ and probabilities $\hat{p_1}, \cdots, \hat{p_n}$. Add jitter $X \sim U(-i/2, i/2)$.

$$E_{p(\theta|y)}[f(\theta)] = \int f(\theta)p(\theta \mid y)d\theta = \int f(\theta)\frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)d\theta}d\theta \approx \frac{\sum_{t=1}^{T}\left[f\left(\theta^{(t)}\right)q\left(\theta^{(t)} \mid y\right)\right]}{\sum_{t=1}^{T} q\left(\theta^{(t)} \mid y\right)}.$$

  Grid sampling **gets computationally too expensive in high dimensions**.

- Rejection Sampling (**important**)

  Target distribution $q(\theta|y)$: hard to sample from. Proposal distribution $g(\theta)$: proxy distribution, easy to sample from.
  We have $q(\theta|y)/Mg(\theta) \leq 1$.

  Rejection Sampling Steps.

  1. Draw a sample $\theta^{(s)}$ from g(θ).
  2. Draw a sample $u^{(s)}$ from $U(0, 1)$.
  3. Compare u_i with $\alpha = q(\theta^{(s)}|y)/Mg(\theta^{(s)})$. Accept if $\mu_i \leq \alpha$.

- Importance Sampling

$$E[f(\theta)] = \int f(\theta)q(\theta)d\theta = \int f(\theta)\frac{q(\theta)}{g(\theta)}g(\theta)d\theta \approx \frac{\sum_s w_s f\left(\theta^{(s)}\right)}{\sum_s w_s}, w_s = \frac{q\left(\theta^{(s)}\right)}{g\left(\theta^{(s)}\right)}.$$

  Draw samples direct from the proposal distribution, then weigh the sample.

2. Markov Chain Monte Carlo(MCMC)

- Markov Chain

  The probability of each event depends only on the state attained in the previous event.

$$K(x, y) = P\left(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0\right) = P\left(X_{t+1} = y \mid X_t = x\right)$$
$$p^{(t+1)}(y) = P(X_{t+1} = y) = \sum_{\cdots} p^{(t)}(x)K(x, y).$$

- Gibbs Sampling

  Stationany Condition. A distribution $\pi(x)$ is stationary with respect to a Markov chain if $\chi^{(t+1)} \sim \pi(x)$ given $X^{(t)} \sim \pi(x)$.
  Reversibility. $\pi(x)K(x, y) = \pi(x)K(y, x)$.
  Gibbs Sampling.

  1. Initialize $(\theta_1, \cdots, \theta_n)$ arbitrarily.
  2. Repeat: Pick j randomly or sequentially. Re-sample θ_j from $p(\theta_j|\theta_{-j})$.(-j means every other term expect the j-th one)

- Metropolis Hasting Algorithm

  Metropolis Algorithm.

1. Starting point θ^0. 2. $t = 1, 2, \cdots,$

2.1 pick a proposal θ^* from the proposal distribution $J_t(\theta^*|\theta^{t-1})$. Proposal distribution has to be symmetric, $i.e.$, $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$.

2.2 Calculate the acceptance ratio

$$r = \frac{p\left(\theta^* \mid y\right)}{p\left(\theta^{t-1} \mid y\right)}.$$

2.3 Set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

Metropolis-Hastings Algorithm.

$$r = \frac{\rho\left(\theta^* \mid y\right) / J_t\left(\theta^* \mid \theta^{t-1}\right)}{\rho\left(\theta^{t-1} \mid y\right) / J_t\left(\theta^{t-1} \mid \theta^*\right)} = \frac{\rho\left(\theta^* \mid y\right) J_t\left(\theta^{t-1} \mid \theta^*\right)}{\rho\left(\theta^{t-1} \mid y\right) J_t\left(\theta^* \mid \theta^{t-1}\right)}.$$

Exercise

[20'] You measure the concentration of N different minerals in a sample of drinking water. Let y_i be a standardized value for the concentration of mineral i. Let $M_i \in \{0, 1\}$ denote whether the levels of mineral i are normal or abnormal. If mineral i is normal $(M_i = 0)$, then $y_i \sim N(0, 1)$. If mineral i is abnormal $(M_i = 1)$ then $y_i = \theta_i + e_i$, where $\theta_i \sim N\left(0, \sigma^2\right)$ and $e_i \sim N(0, 1); \theta_i$ and e_i are independent, and σ^2 is known

(i) What is the marginal distribution for y_i, assuming mineral i is abnormal, $p\left(y_i \mid M_i = 1\right)$? (Hint: Use the property of the sum of two normal variables)

(ii) Let p_1 be the prior probability that mineral i is abnormal, $P\left(M_i = 1\right) = p_1$, for $i = 1, \ldots, N$. Assume the M_i 's are independent given p_1. What is the posterior probability that mineral i is abnormal, $P\left(M_i = 1 \mid y_i, p_1\right)$?

(iii) Assume p_1 is unknown, with prior $p_1 \sim U$ niform $(0, 1)$. What is the posterior distribution of p_1 given the normal/abnormal status of each mineral, $P\left(p_1 \mid M_1, \ldots, M_N\right)$?

(Hint: each M_i can be treated as a bernoulli distribution with parameter p_1) (iv) Describe a Gibbs sampling algorithm to simulate from the joint posterior distribution

$$P\left(M_1, \ldots, M_N, p_1 \mid y_1, \ldots, y_N\right)$$

(Hint: take advantage of the probability we have derived in part ii and iii)

Solution

1.Using the basic property of the sum of two normal variables $(\theta_i + e_i), p(y_i \mid M_i = 1) = N(0, \sigma^2 + 1)$. (This may also be derived directly, by integrating over the prior for θ_i.)

2.

The posterior probability is

$$\frac{p_1 P(y_i \mid M_i = 1)}{p_1 P(y_i \mid M_i = 1) + (1 - p_1) P(y_i \mid M_i = 0)}$$

|where

$$P(y_i \mid M_i = 0) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-y_i^2}{2}\right\} \quad \text{and} \quad P(y_i \mid M_i = 1) = \frac{1}{\sqrt{2\pi(1 + \sigma^2)}} \exp\left\{\frac{-y_i^2}{2(1 + \sigma^2)}\right\}$$

3.

$$P(p_1 \mid M_1, \ldots, M_N) = \text{Beta}\left(1 + \sum_{i=1}^N M_i, 1 + N - \sum_{i=1}^N M_i\right).$$

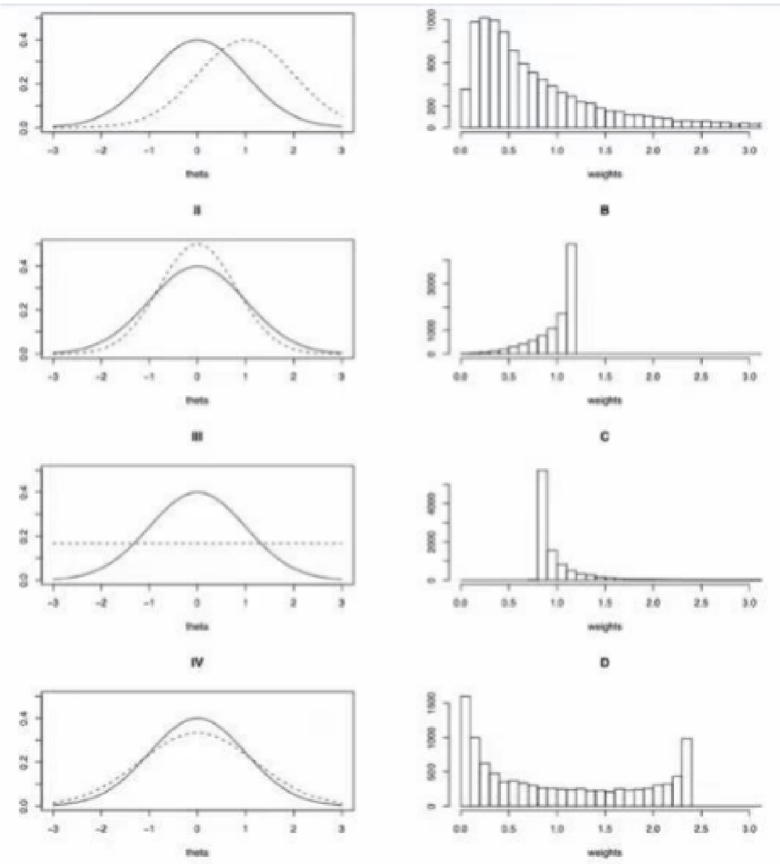
4. $$P(M_1, \ldots, M_N, p_1 \mid y_1, \ldots, y_N).$$

Choose initial value $p_0^{(0)}$. Then, for T samples $t = 1, \ldots, T$:

· Generate $M_i^{(t)}$ from $P(M_i \mid p_1^{(t-1)}, y_1, \ldots, y_n) = p(M_i \mid p_1^{(t-1)}, y_i)$ for $i = 1, \ldots, N$.(Use answer to part b)
· Generate $p_1^{(t)}$ from $P(p_1^{(t)} \mid M_1, \ldots, M_N, y1, \ldots, y_N) = P(p_1^{(t)} \mid M_1, \ldots, M_N)$. (Use answer to part c)

Exercise

Match the following importance sampling figures.



Solution

A,C,D,B

The Asymptotic Theory

- Convergence in distribution

  A sequence of random variables $X_1, X_2, X_3, \cdots$ converges in distribution to random variable X, shown by $X_n \xrightarrow{d} X$, if

  $$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

  for all x at which $F_X(x)$ is continuous.

- Convergence in Probability

  A sequence of random variables $X_1, X_2, X_3, \cdots$ converges in probability to a shown by $X_n \xrightarrow{p} X$, if random variable X,

  $$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0, \quad \text{for all} \epsilon > 0$$

- *Convergence almost surely

Z Let$Z_1, Z_2, \ldots$ be a sequence of rv sin a sample space Ω
in Ω. and Z be another rv in Ω. Then $\{Z_n; n \geq 1\}$ converges converges
if to Z almost surely (a.s.) if

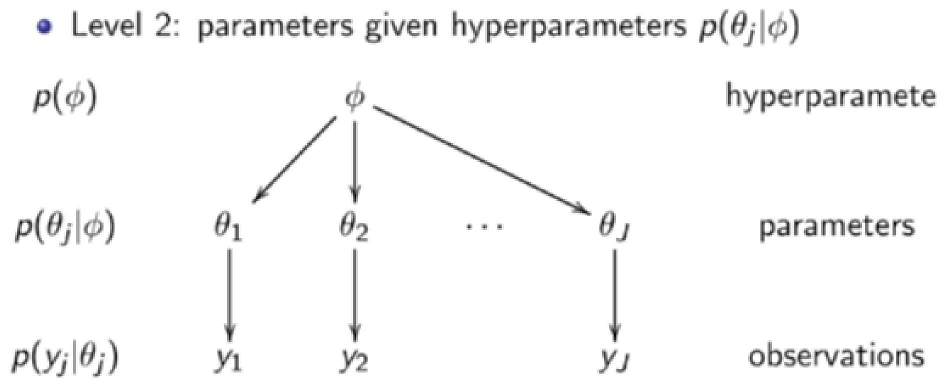
$$\Pr\left\{\omega \in \Omega \colon \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\right\} = 1.$$

Exercise

1. Given that a random process converges almost surely, prove that it must converges in probability
2. Let $\{Y_n; n \geq 1\}$ be a sequence of rv s. If $\sum_{n=1}^{\infty} \mathbb{E}[|Y_n|] < \infty$,then$Y_n \to 0$ i.p (Actually it converges almost surely by Borel Cantelli Lemma)

The Hierarchical Model

Overview



Here we consider three distributions

- Joint Distribution

$$p(\theta, \phi, y) \quad = \quad p(y|\theta, \phi)p(\theta, \phi) \propto \quad p(y|\theta)p(\theta|\phi)p(\phi) = \pi(\phi)\left[\prod_{j=1}^{J} p\left(\theta_j \mid \phi\right) p\left(y_j \mid \theta_j\right)\right]$$

- Conditional Posterior

$$\begin{aligned}
&p(\theta \mid \phi, y) \propto \prod_{j=1}^{J} P\left(\theta_j \mid \phi\right)\mu \\
&P(\theta_1, \cdots, \theta_j \mid \phi, y) \\
&\Rightarrow p\left(\theta_j \mid \phi, y_j\right) \propto p\left(\theta_j \mid \phi\right)
\end{aligned}$$

- Marginal Posterior (Difficult)

The Binomial model

$Joint,\ conditional,\ and\ marginal$ posterior distributions. We first perform the three steps for determining the analytic form of the posterior distribution. The joint posterior distribution of all parameters is

$$\begin{aligned}
p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \\
&\propto p(\alpha, \beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{j}(1-\theta_j)^{n_j-y_j}.
\end{aligned}$$

(5.6)
Given (α, β), the components of θ have independent posterior densities that are of the form $\theta_j^A(1-\theta_j)^B$—that is, beta densities—and the joint density is

$$p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}.$$

(5.7)

We can determine the marginal posterior distribution of (α, β) by substituting (5.6) and (5.7) into the conditional probability formula (5.5):

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}.$$

Exercise

[15'] A childcare chain conducts a study on infectious disease spread across five of their centers. Let y_i be the total count of days missed due to illness over the study period in the i th center. Assume

$$y_i \sim \text{Poisson}(\lambda_i) \text{ and } \lambda_i \sim \text{Gamma}(1, \beta) \text{ for } i = 1, \ldots, 5$$

are mutually independent, unless otherwise specified.
(i) What is the posterior distribution of λ_1, given β and $\mathbf{y}$? i.e. $p(\lambda_1 \mid \beta, y)$
(ii) What is the marginal distribution of y_1, given $\beta, p(y_1 \mid \beta)$?
(iii) What is the posterior distribution for β, given Λ and $y, p(\beta \mid \Lambda, y)$?

Solution

(a) (2 points) What is the posterior distribution of λ_1, given β and $\mathbf{y}$, $p(\lambda_1 \mid \beta, \mathbf{y})$?
By the Poisson-Gamma conjugate model, $\lambda_1 \mid \beta, y_1 \sim \textit{Gamma}(1 + y_1, \beta + 1)$.

(b) (2 points) What is the marginal distribution of y_1, given β, $P(y_1 \mid \beta)$?
This is the marginal distribution for a Poisson-Gamma model, which we have shown in clas is Negative Binomial (NB). Here is the derivation:

$$\begin{aligned}
p(y_1 \mid \beta) &= \int_0^\infty P(y_1 \mid \lambda_1) p(\lambda_1 \mid \beta) d\lambda_1 \\
&= \int_0^\infty \frac{e^{-\lambda_1} \lambda_1^{y_1}}{y_1!} \beta e^{-\beta\lambda_1} d\lambda_1 \\
&= \frac{\beta}{y_1!} \int_0^\infty \lambda_1^{y_1} e^{-\lambda_1(\beta+1)} d\lambda_1 \\
&= \frac{\beta}{y_1!} \frac{\Gamma(y_1+1)}{(\beta+1)^{y_1+1}} \\
&= \frac{\beta}{(\beta+1)^{y_1+1}}
\end{aligned}$$

Which is an $NB\left(1, \frac{\beta}{\beta+1}\right)$ *distribution.*

(c) (3 points) What is the posterior distribution for β, given $\mathbf{\Lambda}$ and $\mathbf{y}$, $p(\beta \mid \mathbf{\Lambda}, \mathbf{y})$?
Note that

$$\begin{aligned}
p(\beta \mid \mathbf{\Lambda}) &\propto p(\beta) \prod_{i=1}^{5} p(\lambda_i \mid \beta) \\
&\propto e^{-\beta} \prod_{i=1}^{5} \beta e^{-\beta\lambda_i} \\
&= \beta^5 e^{-\beta(1+\sum_{i=1}^5 \lambda_i)}
\end{aligned}$$

which is the kernel for a $\textit{Gamma}\left(6, 1 + \sum_{i=1}^5 \lambda_i\right)$ *distribution.*

Exchangeability

The set $Y_1, Y_2, \ldots, Y_n$ is exchangeable if the joint probability $p(y_1, \ldots, y_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p(y_1, \ldots, y_n) = p(y\pi_1, \ldots, y\pi_n).$$

The set $Y_1, Y_2, \ldots, Y_n$ is infinitely exchangeable if, for any n,$Y_1, Y_2, \ldots, Y_n$are exchangeable.