Simulación ITAM
8 de noviembre de 2020 2020-II

Tarea 5.

No hay fecha de entrega, son problemas de práctica.Yo subiré soluciones el 1 de diciembre.

Lecturas

- Robert & Casella Capítulos 6 y 7
- Dagpunar Capítulos 5 y 8
- Efron y Gong: A leisurely look at the Boostrap, the Jackknife and Cross-Validation
- Chib y Greenberg: Understanding the Metropolis-Hastings Algorithm
- Casella & George. Explaining the Gibbs sampler.

Problemas

1. Supongan que $Y|\theta \sim \mathcal{G}(1, \theta)$ y que $\theta \sim IG(\alpha, \beta)$.
   - Encuentren la distribución posterior de θ.
   - Encuentren la media y varianza posterior de θ.
   - Encuentren la moda posterior de θ.
   - Escriban dos ecuaciones integrales que se pueden resolver para encontrar el intervalo de 95 % de colas simétricas para θ
2. Los siguientes datos corresponden a las horas adicionales de sueño de 10 pacientes tratados con un somnífero B comparado con un somnífero A:

   $$1.2, 2.4, 1.3, 1.3, 0, 1, 1.8, 0.8, 4.6, 1.4$$

   Lleven a cabo un análisis bayesiano de estos datos y extraigan conclusiones, asumiendo cada componente de la verosimilitud que sea:
   - normal
   - $t_{(3)}$
   - $t_{(1)}$

- Bernoulli (de alguna manera que se les ocurra)

En este ejercicio, escriban un código para manejar cualquier integración necesaria y cálculo de probabilidades marginales posteriores.

3. Spiegelhalter et al. (1995) analiza la mortalidad del escarabajo del trigo en la siguiente tabla, usando BUGS.

Dosis	# muertos	# expuestos
w_i	y_i	n_i
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Estos autores usaron una parametrización usual en dos parámetros de la forma $p_i \equiv P(\text{muerte}|w_i)$, pero comparan tres funciones ligas diferentes:

$$\begin{aligned}\text{logit: } p_i &= \frac{\exp(\alpha+\beta z_i)}{1+\exp(\alpha+\beta z_i)}\\ \text{probit: } p_i &= \Phi(\alpha+\beta z_i)\\ \text{complementario log-log: } p_i &= 1-\exp[-\exp(\alpha+\beta z_i)]\end{aligned}$$

en donde se usa la covariada centrada $z_i = w_i - \bar{w}$ para reducir la correlación entre la ordenada α y la pendiente β. En OpenBUGS el código para implementar este modelo es el que sigue:

```
model{
  for (i in 1:k){
    y[i] ~ dbin(p[i],n[i])
        logit(p[i]) <- alpha + beta*(w[i]-mean(w[])
        #       probit(p[i]) <- alpha + beta*(w[i]-mean(w[])
        #       cloglog(p[i]) <- alpha + beta*(w[i]-mean(w[])
      } #fin del loop i

      alpha ~ dnorm(0.0,1.0e-3)
      dbeta ~ dnorm(0.0,1.0e-3)
} #fin del código
```

Lo que sigue al símbolo # es un comentario, así que esta versión corresponde al modelo logit. También `dbin` denota la distribución binomial y `dnorm` denota la distribución normal, donde el segundo argumento denota la precisión, no la varianza (entonces las iniciales normales para α y β tienen precisión 0.001, que son aproximadamente iniciales planas (no informativas)). Hacer el análisis en `OpenBUGS`.

4. Consideren las siguientes dos distribuciones condicionales completas, analizadas en el artículo de Casella y George (1992) que les incluí como lectura:

$$\begin{aligned} f(x|y) &\propto ye^{-yx}, \qquad 0 < x < B < \infty \\ f(y|x) &\propto xe^{-xy}, \qquad 0 < y < B < \infty \end{aligned}$$

   - Obtener un estimado de la distribución marginal de X cuando $B = 10$ usando el Gibbs sampler.
   - Ahora supongan que $B = \infty$ así que ls distribuciones condicionales completas son ahora las ordinarias distribuciones exponenciales no truncadas. Mostrar analíticamente que $f_x(t) = 1/t$ es una solución a la ecuación integral en este caso:

   $$f_x(x) = \int \left[\int f_{x|y}(x|y) f_{y|t}(y|t) dy \right] f_x(t) dt$$

   ¿El Gibbs sampler convergerá a esta solución?

5. En una prueba real, 12 lotes de mantequilla de cacahuate tienen residuos de aflatoxin en partes por mil millones de 4.94, 5.06, 4.53, 5.07, 4.99, 5.16, 4.38, 4.43, 4.93, 4.72, 4.92, y 4.96.

   - ¿Cuántas posibles muestras bootstrap hay en estos datos?
   - Usando `R` y la función `sample`, o una tabla de números aleatorios, generar 100 remuestras de los datos de la muestra. Para cada una de estas remuestras, obtener la media. Comparar la media de las medias obtenidas en las remuestras con la media de la muestra original.
   - Encontrar de las 100 remuestras, un intervalo de confianza del 95 % para la media.

6. El número de accidentes aéreos de 1983 a 2006 fueron 23, 16, 21, 24, 34, 30, 28, 24, 26, 18, 23, 23, 36, 37, 49, 50, 51, 56, 46, 41, 54, 30, 40, 31.

   - Para la muestra de datos, calcular la media y su error estándar (a partir de la desviación estándar), así como la mediana.
   - Usando `R`, calcular estimados bootstraps de la media y la mediana con estimados de sus errores estándar, usando $B = 1000$ remuestras. También calcular la mediana de las medianas muestrales.
   - ¿Cómo se comparan los dos incisos anteriores?

7. Supongan que una variable aleatoria y se distribuye de acuerdo a la densidad poli-Cauchy:

$$g(y) = \prod_{i=1}^{n} \frac{1}{\pi(1 + (y - a_i)^2)}$$

donde $a = (a_1, \ldots a_n)$ es un vector de parámetros. Supongan que $n = 6$ y $a = (1, 2, 2, 6, 7, 8)$.

   - Escriban una función que calcule la log-densidad de y.
   - Escriban una función que tome una muestra de tamaño 10,000 de la densidad de y, usando Metropolis-Hastings con función propuesta una caminata aleatoria con desviación estandar C. Investiguen el efecto de la elección de C en la tasa de aceptación, y la mezcla de la cadena en la densidad.
   - Usando la muestra simulada de una "buena" elección de C, aproximar la probabilidad $P(6 < Y < 8)$.

8. Supongan que el vector (X, Y) tiene función de distribución conjunta:

$$f(x, y) = \frac{x^{a+y-1} e^{-(1+b)x} b^a}{y! \Gamma(a)}, x > 0, y = 0, 1, 2, \cdots$$

y deseamos simular de la densidad conjunta.

   - Mostrar que la densidad condicional $f(x|y)$ es una Gamma e identificar los parámetros.
   - Mostrar que la densidad condicional $f(y|x)$ es Poisson.
   - Escriban una función para implementar el Gibbs sampler cuando las constantes son dadas con valores $a = 1$ y $b = 1$.
   - Con su función, escriban 1000 ciclos del Gibss sampler y de la salida, hacer los historgramas y estimar $E(Y)$.

9. La τ de Kendall entre X y Y es 0.55. Tanto X como Y son positivas. ¿Cuál es la τ entre X y $1/Y$? ¿Cuál es la τ de $1/X$ y $1/Y$?

10. Mostrar que cuando $\theta \to \infty$, $C^{Fr}(u_1, u_2) \to \text{mín}\{u_1, u_2\}$, donde C^{Fr} es la cópula de Frank.

11. Consideren la cópula de Clayton. Mostrar que converge a la cópula de comonotonicidad cuando $\theta \to \infty$. [Hint: usen la regla de l'Hópital considerando que la cópula de Clayton se puede escribir como $\exp\{log(u_1^{-\theta} + u_2^{-\theta} - 1)/\theta\}$ para θ positivo.]

12. Supongan que tienen dos vectores de datos $(x_1, \ldots, x_n)$ y $(y_1, \ldots, y_n)$. Entonces la cópula empírica es la función $C : [0, 1] \times [0, 1] \to [0, 1]$ definida por

$$C(u, v) = \frac{1}{n} \sum_{j=1}^{n} I\left(\frac{r_j}{n+1} \leq u, \frac{s_j}{n+1} \leq v\right)$$

donde $(r_1, \dots, r_n)$ y $(s_1, \dots, s_n)$ denotan los vectores de rangos de x y y respectivamente.

Escriban una función llamada `empCopula` que tome cuatro argumentos `u, v, xVec` y `yVec`. Pueden suponer que los valores `u,v` están en $[0, 1]$ y que `xVec` y `yVec` son vectores numéricos que tienen la misma longitud (no vacios).

13. la cópula Farlie-Gumbel-Morgenstern es $C(u, v) = uv[1+\alpha(1-u)(1-v)]$ para $|\alpha| \leq 1$. Mostrar que la densidad conjunta correspondiente $\frac{\partial^2 C(u,v)}{\partial u \partial v}$ es no negativa. Mostrar que C tiene marginales uniformes en $(0, 1)$. Encontrar el coeficiente de correlación de Spearman y la tau de Kendall.

14. Este es un ejercicio de calibración de las cópulas utilizando correlaciones de rangos. Supongan que una muestra produce un estimado de la τ de Kendall de 0.2. ¿Qué parámetro debe usarse para

    i. la cópula normal,

    ii. la cópula de Gumbel,

    iii. la cópula de Clayton?

15. Usen la función `normalCopula` del paquete `copula` para crear una cópula gaussiana bidimensional con un parámetro de 0.9. Luego creen otra cópula gaussiana con parámetro de 0.2 y describan la estructura de ambas cópulas (diferencias y semejanzas).

16. Usen la función `rCopula` del paquete `copula` para generar muestras de 500 puntos cuya distribución son las cópulas del ejercicio 8 anterior. Hagan una gráfica de las dos muestras. Teniendo en mente que una cópula determina la estructura de dependencia de una distribución multivariada conjunta, mirando estas gráficas, ¿pueden decir cuál de estas dos cópulas debe usarse para simular una distribución con una fuerte dependencia entre las marginales?