CAPÍTULO 8. INTERCAMBIABILIDAD Y MODELOS JERÁRQUICOS

Para leer

Bernardo y Smith (1994), Secciones 4.2 y 4.3.

Gelman et al (1995), Capítulo 5,

Lee (1997), Capítulo 8.

En este Capítulo se consideran las situaciones en las que existe una estructura en el modelo que se quiere incluir en la formulación de la distribución a priori.

Intercambiabilidad

La intercambiabilidad (De Finetti 1964, 1974) proporciona una idea más débil de simetría de la distribución.

Definición 13 *Un conjunto de sucesos aleatorios* $X_1, \ldots, X_n$ *es* **intercambiable** *si cualquier permutación tiene la misma distribución que cualquier otra permutación, es decir que si* j_1, $\ldots, j_m$ *son números enteros en* $1, \ldots, n$, *entonces*

$$P(\cap_{i=1}^{m} X_{j_i}) = P(\cap_{i=1}^{m} X_i)$$

Las variables aleatorias $X_1, X_2, \ldots$ *son (ínfinitamente) intercambiables si para* m *números enteros,* $j_1, \ldots, j_m$, *tenemos*

$$(X_{j_1}, \ldots, X_{j_m}) \sim (X_1, \ldots, X_m)$$

donde $m \in 1, 2, \ldots$

Ejemplo 100 *Si $X_1, \ldots, X_n$ son i.i.d., son intercambiables.*

Ejemplo 101 *Si $X_i =$ carta i es una espada, los X_i son intercambiables pero no son independientes.*

Ejemplos de modelos no intercambiables son los modelos estocásticos (cadenas de Markov, colas ...).

Teoremas de De Finetti

Hay dos resultados que clasifican series intercambiables. El primer resultado (De Finetti 1930, 1964) es para variables Bernoulli.

Teorema 16 *Si $X_1, X_2, \ldots$ es una sucesión ínfinitamente intercambiable de variables 0-1 con medida de probabilidad $P(\cdot)$ entonces existe una función de distribución $F(\cdot)$, tal que,*

$$p(x_1, \ldots, x_m) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{m-x_i}\, dF(\theta)$$

donde $F(\theta) = \lim_{n\to\infty} P(Y_n/n \leq \theta)$, $Y_n = \sum_{i=1}^{n} X_i$ y $\lim_{n\to\infty} Y_n/n = \theta$.

Demostración

Ver Bernardo y Smith (1994).

Interpretación

El resultado implica que suponiendo la intercambiabilidad ínfinita, existe θ tal que las $X_i|\theta$ sean variables i.i.d. Bernoulli donde el parámetro θ tiene una distribución a priori $F(\theta)$. Podemos pensar que el contenido de $F(\cdot)$ es una representación de nuestras creyencias sobre la frecuencia relativa de 1's en un número grande de observaciones.

El teorema general

Teorema 17 *Sea $X_1, X_2, \ldots$ una sucesión de variables ínfinitamente intercambiables. Defina $P(x) = P(X \leq x)$. Entonces, existe una distribución Q tal que*

$$P(x_1, \ldots, x_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(x_i) dQ(F)$$

donde $\mathcal{F}$ es el espacio de funciones de distribución, $Q(F) = \lim_{n\to\infty} P(F_n)$ y F_n es la función de distribución empírica definida por $\mathbf{x}$.

El teorema implica que si los X_i son intercambiables, existe $\boldsymbol{\theta}$ para que

$$P(\mathbf{x}) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) dQ(\boldsymbol{\theta}).$$

Observación 63 *Los teoremas de De Finetti justifican la inferencia bayesiana teóricamente, pero no son muy útiles para determinar la distribución a priori en la práctica.*

Algunas extensiones justifican modelos específicos. Por ejemplo se dice que $X_1, \ldots$ *presenta simetría esférica si* $\mathbf{X} \sim \mathbf{AX}$ *para* $\mathbf{A}$ *ortogonal.*

Los supuestos de intercambiabilidad ínfinita y simetría esférica justifican un modelo normal dada alguna distribución a priori para la media y varianza.

Ver Bernardo y Smith (1994).

Módelos jerárquicos

Supongamos que se hace una serie de experimentos donde los datos $\mathbf{Y}_j$ del experimento i dependen de parámetros θ_j, para $j = 1, \ldots, J$.

Típicamente se hacen varios suposiciones sobre la relación entre los parámetros θ_j. Por ejemplo que son i.i.d. o que $\theta_1 < \ldots < \theta_J$. Tales relaciones se llaman estructurales.

En muchos casos, la información estructural se combina con una distribución a priori sobre los parámetros de la estructura o híperparametros.

En este caso, se dice que la distribución a priori es **jerárquica**. Ver Good (1980).

Ejemplo 102 *Varios individuos hacen una prueba de inteligencia donde se supone que sus resultados tienen una distribución normal*

$$Y_i|\theta_i, \phi \sim \mathcal{N}\left(\theta_i, \frac{1}{\phi}\right)$$

donde la media depende de la inteligencia real del sujeto. Se puede suponer también que la población de medias es normal, es decir que

$$\theta_i|\mu, \psi \sim \mathcal{N}\left(\mu, \frac{1}{\psi}\right)$$

donde los híperparámetros son μ y ψ.

Ejemplo 103 *George et al (1993) analizaron datos de fallos de bombas eléctricas. Se suponian que si X_i es el número de fallos en la bomba i, entonces*

$$X_i|\theta_i, t_i \sim \mathcal{P}(t_i\theta_i) \quad \textit{para } i = 1, \ldots, 10$$

donde t_i es el tiempo de funcionamiento de la bomba.

Se empleó una distribución a priori jerárquica:

$$\begin{aligned} \theta_i|\alpha, \beta &\sim \mathcal{G}(\alpha, \beta) \\ \alpha &\sim \mathcal{E}(1) \\ \beta &\sim \mathcal{G}(0{,}1, 1) \end{aligned}$$

La estrategia general

Consideramos un conjunto de experimentos $j = 1, \ldots, J$ en el que el experimento j-ésimo proporciona datos $\mathbf{x}_j$, con parámetro $\boldsymbol{\theta}_j$ y verosimilitud $l(\boldsymbol{\theta}_j|\mathbf{x}_j)$. Si no hay otra información salvo los datos para distinguir los parámetros $\boldsymbol{\theta}_j$, se puede representar tal simetría mediante intercambiabilidad de los parámetros. Luego

$$f(\boldsymbol{\theta}|\boldsymbol{\phi}) = \prod_{j=1}^{J} f(\boldsymbol{\theta}_j|\boldsymbol{\phi})$$

con la incertidumbre sobre $\boldsymbol{\phi}$ modelizada usando una hiperdistribución $f(\boldsymbol{\phi})$.

Entonces, la distribución conjunta a priori es

$$f(\boldsymbol{\theta}, \boldsymbol{\phi}) = f(\boldsymbol{\theta}|\boldsymbol{\phi}) f(\boldsymbol{\phi})$$

y la distribución conjunta a posteriori es

$$f(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{x}) \propto f(\boldsymbol{\theta}, \boldsymbol{\phi}) l(\mathbf{x}|\boldsymbol{\theta})$$

Además, la densidad predictiva es

$$f(x|\mathbf{x}) = \int \int f(x|\boldsymbol{\theta}, \boldsymbol{\phi}) f(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) f(\boldsymbol{\phi}|\mathbf{x}) \, d\boldsymbol{\phi} \, d\boldsymbol{\theta}$$

Un problema es que se tiene que asignar una distribución a priori sobre $\boldsymbol{\phi}$.

Exploramos varias posibilidades.

Distribución a priori propia

Si se impone una distribución a priori propia $f(\boldsymbol{\phi})$ luego, se emplean los métodos de muestreo Gibbs para hallar la distribución a posteriori de $\boldsymbol{\theta}$ porque en muchas situaciones, las distribuciones

$$f(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\phi}) \; y \; f(\boldsymbol{\phi}|\mathbf{x}, \boldsymbol{\theta}).$$

pueden ser fáciles de hallar.

No obstante, el problema es que a menudo no existe mucha información real sobre los hiperparametros.

El método Bayesiano empírico

En este caso, se sustituyen los híperparámetros ϕ por algún valor estimado $\hat{\phi}$ y se procede con el análisis bayesiano habitual, suponiendo que los hiperparametros son conocidos.

Ver, por ejemplo, Carlin y Louis (1996, 2000).

Ejemplo 104 *Sea*

$$\begin{aligned} X_i|\theta_i &\sim \mathcal{N}(\theta_i, 1) \\ \theta_i|\tau^2 &\sim \mathcal{N}(0, \tau^2) \end{aligned}$$

para $i = 1, \ldots, n$.

No se quiere especificar una distribución a priori para τ^2. *¿Cómo se puede estimar?*

La distribución marginal de $\mathbf{X}$ *es*

$$\mathbf{X}|\tau^2 \sim \mathcal{N}\left(\mathbf{0}, (1+\tau^2)\mathbf{I}_n\right)$$

y dados los datos, el EMV de τ^2 *es*

$$\hat{\tau}^2 = \text{máx}\left\{0, \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1\right\}.$$

Ahora, se procede suponiendo como si τ^2 *fuera conocido. Cuando* $\hat{\tau}^2 > 0$*, se tiene*

$$\theta_i|\mathbf{x} \sim \mathcal{N}\left(\frac{\widehat{\tau^2}}{1+\widehat{\tau^2}}x_i, \frac{\widehat{\tau^2}}{1+\widehat{\tau^2}}\right)$$

y entonces

$$E[\boldsymbol{\theta}|\mathbf{x}] = \left(1 - \frac{n}{\sum_i x_i^2}\right)^+ \mathbf{x}.$$

Observación 64 *Este estimador es un estimador de Stein truncado. Ver Stein (1955), James y Stein (1960).*

Un problema con este método es que es muy arbitrario ya que se pueden elegir varias estimadores de ϕ (EMV, método de momentos etc.) Otro problema es que no es muy bayesiano.

Utilizar distribuciones a priori no informativos

Ejemplo 105 *Retomando el Ejemplo 102, supongamos en principio que se conocen los valores de ϕ y ψ y que se impone una distribución uniforme a priori para μ.*

Luego

$$f(\mu, \boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{\phi}{2}\sum_i (y_i - \theta_i)^2\right) \exp\left(-\frac{\psi}{2}\sum_i (\theta_i - \mu)^2\right).$$

Integrando con respeto a μ, se tiene

$$\begin{aligned} \boldsymbol{\theta}|\mathbf{y} &\sim \mathcal{N}(\mathbf{m}, \mathbf{W}) \quad \textit{donde} \\ \mathbf{W}^{-1} &= \frac{1}{\phi+\psi}\mathbf{I} + \frac{\psi}{n\phi(\phi+\psi)}\mathbf{J} \\ \mathbf{W}\mathbf{m} &= \phi\mathbf{y} \end{aligned}$$

y la media a posteriori de $\boldsymbol{\theta}$ es

$$E[\boldsymbol{\theta}|\mathbf{y}]0\frac{\phi}{\phi+\psi}\mathbf{y} + \frac{\psi}{\psi+\phi}\mathbf{1}\bar{y}$$

es decir que la media de θ_j es una media ponderada del EMV y la media global $\bar{y}$.

No obstante si se impone una distribución impropia a priori para ϕ y ψ, es posible demostrar que la distribución a posteriori es impropia.

CAPÍTULO 9. MÉTODOS NUMÉRICOS

Para leer

Gelman et al Capítulo 9, Seccion 9.5, Capítulo 10.

Dada una distribución a priori $f(\boldsymbol{\theta})$ y unos datos $\mathbf{x}$, a menudo se quiere evaluar:

- la densidad a posteriori

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{x})}{\int f(\boldsymbol{\theta})l(\boldsymbol{\theta}|\mathbf{x})\ d\boldsymbol{\theta}}$$

- suponiendo una pérdida $L(\mathbf{T}, \boldsymbol{\theta})$, el estimador Bayes

$$\min_{\mathbf{T}} E[L(\mathbf{T}, \boldsymbol{\theta})|\mathbf{x}]$$

- intervalos de credibilidad, ...

Excepto las pocas situaciones en que tenemos familias conjugadas, el cálculo de las integrales es problematica. Se necesitan emplear métodos numéricos.

Ejemplo 106 $X|\theta \sim \mathcal{C}(1,\theta)$. *Dada una distribución a priori no-informativa,* $f(\theta) \propto 1$, *la distribución a posteriori es*

$$f(\theta|\mathbf{x}) \propto \prod \left[1 + (x_i - \theta)^2\right]^{-1}$$

y no se puede calcular ni la constante de integración ni los momentos sin métodos aproximados.

Observación 65 *Como la distribución a posteriori va a ser bastante simétrica, si la muestra es suficientemente grande, una posibilidad es emplear una aproximación normal, como se ha visto en el Capítulo 6.*

Ejemplo 107 *Sea*

$$f(x|\mathbf{w},\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i g(x|\theta_i)$$

una mixtura de densidades.

Dados los datos, la verosimilitud es

$$l(\mathbf{w},\boldsymbol{\theta}|\mathbf{x}) \propto \prod_{j=1}^{n} \sum_{i=1}^{k} w_i g(x_j|\theta_i)$$

que tiene k^n *términos.*

Para n *(y* k*) grande, no es posible evaluar la distribución a posteriori de* $(\mathbf{w},\boldsymbol{\theta})$ *en un tiempo finito.*

Observación 66 *Si θ es discreto (y infinito), a menudo hay menos dificultades en evaluar la densidad porque estimar una suma es más fácil que estimar una integral. Se trunca la suma para estimar la constante de la distribución.*

$$P(\theta|\mathbf{x}) \approx \frac{P(\theta)l(\theta|\mathbf{x})}{\sum_i^N P(\theta = i)l(\theta = i|\mathbf{x})}$$

donde N es un número grande. Utilizando valores diferentes de N se puede determinar una aproximación suficientemente precisa.

Ejemplo 108 *Sea $X|N \sim \mathcal{BI}(N, 1/2)$ y la distribución a priori geométrica $N - 1 \sim \mathcal{GE}(1/4)$:*

$$P(N = n) = \frac{1}{4}\left(\frac{3}{4}\right)^{n-1} \qquad n = 1, 2, \ldots$$

Dada una observación $x = 5$ la distribución a posteriori es

$$\begin{aligned} P(N = n|x = 5) &\propto \left(\frac{3}{4}\right)^{n-1} \binom{n}{5} \frac{1}{2^n} \\ &\propto \frac{n!}{(n-5)!}\left(\frac{3}{8}\right)^n \quad \text{para } n \geq 5. \end{aligned}$$

No es una distribución estandar y debemos aproximar su constante mediante truncación.

la tabla muestra los valores estimados de la constante y de la media a posteriori truncando la suma en varios valores de N.

$N_{\text{máx}}$	$\sum_{n=5}^{N_{\text{máx}}} \frac{n!}{(n-5)!}\left(\frac{3}{8}\right)^n$	$E[N\|x=5]$
10	12,028	7,688
15	14,763	8,504
20	14,925	8,595
40	14,930	8,600
60	14,930	8,600

Se ve que truncando el sumatorio en $N_{\text{máx}} = 40$, *se tiene convergencia suficiente para aproximar la media en dos decimales.*

Observación 67 *Volviendo al Ejemplo 107, se tienen los mismos problemas si* $\boldsymbol{\theta}$ *es discreta o continua, sea conocida o no* $\mathbf{w}$.

Integración Númerica

Existen muchos métodos.

La regla de Simpson

Se quiere evaluar la integral

$$I = \int_a^b g(\theta)\, d\theta$$

En su versión más simple, la regla de Simpson es

$$I \approx \frac{b-a}{6}\left[g(a) + 4g\left(\frac{a+b}{2}\right) + g(b)\right].$$

Se mejora la aproximación dividiendo el intervalo $[a, b]$ en un número par N de subintervalos

$$[a, a+h) \cup \cdots \cup [a+(N-1)h, a+Nh = b]$$

Utilizando la regla de Simpson para cada subintervalo, se halla el estimador

$$\begin{aligned} I \approx & \frac{h}{3}[g(a)+4g(a+h)+2g(a+2h)+\cdots \\ & +2g(a+(N-2)h)+4g(a+(N-1)h)+ \\ & +g(a+Nh)] \end{aligned}$$

Ejemplo 109 *Supongamos que* $\theta|\mathbf{x} \sim \mathcal{B}(7,10)$. *Entonces*

$$f(\theta|\mathbf{x}) \propto \theta^6(1-\theta)^9$$

Intentamos estimar

$$I = \int_0^1 \theta^6(1-\theta)^9 d\theta$$

Usando la regla de Simpson con $h = 0{,}1$ *se obtiene la siguiente tabla:*

θ	$\theta^6(1-\theta)^9$	$\int_0^\theta \phi^6(1-\phi)^9\,d\phi$
,0	$0{,}00000E-00$	$0{,}00000E-00$
,1	$3{,}87420E-07$	
,2	$8{,}58994E-06$	$3{,}37987E-07$
,3	$2{,}94178E-05$	
,4	$4{,}12783E-05$	$5{,}92263E-06$
,5	$3{,}05176E-05$	
,6	$1{,}22306E-05$	$1{,}17753E-05$
,7	$2{,}31568E-06$	
,8	$1{,}34218E-07$	$1{,}24962E-05$
,9	$5{,}31441E-10$	
1,0	$0{,}00000E-00$	$1{,}25007E-05$

El valor exacto de la integral es $1{,}24875E-05$. *(La regla de Simpson con* $h = 0{,}05$ *nos lleva el resultado* $1{,}24876E-05$.*)*

Se pueden añadir más columnas a la tabla para estimar, por ejemplo, la media.

Ejemplo 110 *Se muestrean 5 observaciones de una distribución Cauchy con densidad*

$$f(x|\theta) = \frac{1}{\pi\left\{1+(x-\theta)^2\right\}}.$$

Los datos son $\mathbf{x} = (11{,}4, 7{,}3, 9{,}8, 13{,}7, 10{,}6)^T$. *Se supone una distribución a priori uniforme para* θ.

Supongamos que queremos estimar la constante de integración, $P(\theta < 11{,}5|\mathbf{x})$ *y la media a posteriori de* θ.

La densidad a posteriori es

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \left[1+(11{,}4-\theta)^2\right]\ldots\left[1+(10{,}6-\theta)^2\right] \\ &\propto H(\theta) \qquad \textit{donde} \end{aligned}$$

$$H(\theta) = 10^5\left[1+(11{,}4-\theta)^2\right]\ldots\left[1+(10{,}6-\theta)^2\right].$$

Observación 68 *Es útil multiplicar por una constante* (10^5) *para disminuir los problemas números.*

Para calcular la constante, es preciso estimar $\int_{-\infty}^{\infty} H(\theta)\ d\theta$. *Utilizamos la regla de Simpson con* $h = 0{,}5$.

θ	$H(\theta)$	$\theta H(\theta)$	$\int_{-\infty}^{\theta} H(t)\ dt$	$F(\theta\|\mathbf{x})$	$\int_{-\infty}^{\theta} tH(t)\ dt$
4,5	,0	,0	,0	,000	,0
5,0	,0	,0			
5,5	,0	,1	,0	,000	,0
6,0	,1	,4			
6,5	,2	1,4	,1	,000	0,5
7,0	,8	5,6			
7,5	2,3	16,9	1,0	,002	7,3
8,0	4,8	38,8			
8,5	10,7	90,7	6,4	,012	51,1
9,0	28,2	253,9			
9,5	82,7	785,9	40,8	,076	366,4
10,0	196,1	1961,4			
10,5	290,6	3051,2	233,8	,436	2313,5
11,0	250,1	2751,1			
11,5	129,2	1485,6	470,5	,877	4903,8
12,0	47,4	568,2			
12,5	17,3	216,3	526,4	,982	5566,3
13,0	7,4	96,3			
13,5	3,3	44,0	534,8	,997	5673,9
14,0	1,1	15,4			
14,5	,3	4,2	536,1	1,000	5692,2
15,0	,1	1,2			
15,5	,0	,4	536,2	1,000	5693,7
16,0	,0	,1			
16,5	,0	,0	536,3	1,000	5693,8

Observación 69 *Utilizamos la aproximación*

$$\int_{-\infty}^{\infty} H(\theta)\, d\theta \approx \int_{4,5}^{16,5} H(\theta)\, d\theta$$

porque la densidad es aproximadamente cero fuera del intervalo $[4{,}5, 16{,}5]$.

$P(\theta < 11{,}5|\mathbf{x}) \approx 470{,}5/536{,}3 = {,}877$. *Se estima la media a posteriori con*

$$E[\theta|\mathbf{x}] \approx 5693{,}8/536{,}3 \approx 10{,}62$$

Observación 70 *Se puede estimar cualquiera integral utilizando la regla de Simpson pero no es un método muy preciso. Existen problemas en casos multidimensionales.*

Cuadratura de Gauss-Hermite

Se utiliza la aproximación

$$\int_{-\infty}^{\infty} e^{-t^2/2} g(t)\, dt \approx \sum_{i=1}^{N} w_i f(t_i)$$

donde los t_i son los ceros del polinomio de Hermite $H_N(t)$ y

$$w_i = \frac{2^{N-1} N! \sqrt{N}}{N^2 (H_{N-1}(t_i))^2}$$

Observación 71 *Si la distribución se parece a una normal, el método es bastante eficiente. Se puede extenderlo al caso de $\boldsymbol{\theta}$ multidimensional pero la eficacia disminuye con dim($\boldsymbol{\theta}$). Existen otras aproximaciones gaussianas para usar cuando el soporte es finito.*

Otros métodos de integración numérica: Romberg (trapezoidal), polinomiales, etc. Ver Evans (1993).

Métodos Monte-Carlo

Ver, por ejemplo Chen et al (2000).

Se quiere evaluar la integral

$$E[g(\theta)|\mathbf{x}] = \int g(\theta) f(\theta|\mathbf{x})\, d\theta$$

Si se puede muestrear $f(\theta|\mathbf{x})$, entonces, se estima $E[g(\theta)|\mathbf{x}]$ con

$$\widehat{g} = \frac{1}{N} \sum_{i=1}^{N} g(\theta^{(i)})$$

donde $\theta^{(1)}, \ldots, \theta^{(N)}$ son los datos muestreados. El estimador tiene varias propiedades buenas.

Propiedades

- cuando $N \to \infty$, $\widehat{g} \to E[g(\theta)|\mathbf{x}]$.

- mediante el teorema de limite central, un intervalo (aproximado) de confianza de 95 % para $E[g(\theta)|\mathbf{x}]$ es

  $$\widehat{g} \pm 2\widehat{se}(\widehat{g}).$$

  El tamaño del intervalo nos proporciona una idea de la precisión de nuestra estimación. Si es demasiado ancho, siempre podemos muestrear más datos.

- Se necesita suponer la existencia de $V[g(\theta)|\mathbf{x}]$.

Ejemplo 111 *Sea* $X|\mu,\phi \sim \mathcal{N}(\mu, 1/\phi)$ *con distribución a priori de Jeffreys* $f(\mu,\phi) \propto 1/\phi$. *Entonces*

$$\begin{aligned} \phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \\ \mu|\phi &\sim \mathcal{N}(\bar{x}, 1/(n\phi)) \end{aligned}$$

Se puede muestrear $f(\mu,\phi|\mathbf{x})$ *mediante el siguiente algoritmo*

1. $i = 1$

2. Generar $\phi^{(i)}$ de $\mathcal{G}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$,

3. Generar $\mu^{(i)}$ de $\mathcal{N}\left(\bar{x}, 1/(\phi^{(i)}n)\right)$

4. $i = i + 1$. Si $i \leq N$, Goto 2.

Se estima, por ejemplo, la media $E[\mu|\mathbf{x}]$ *con* $\frac{1}{N}\sum \mu^{(i)}$.

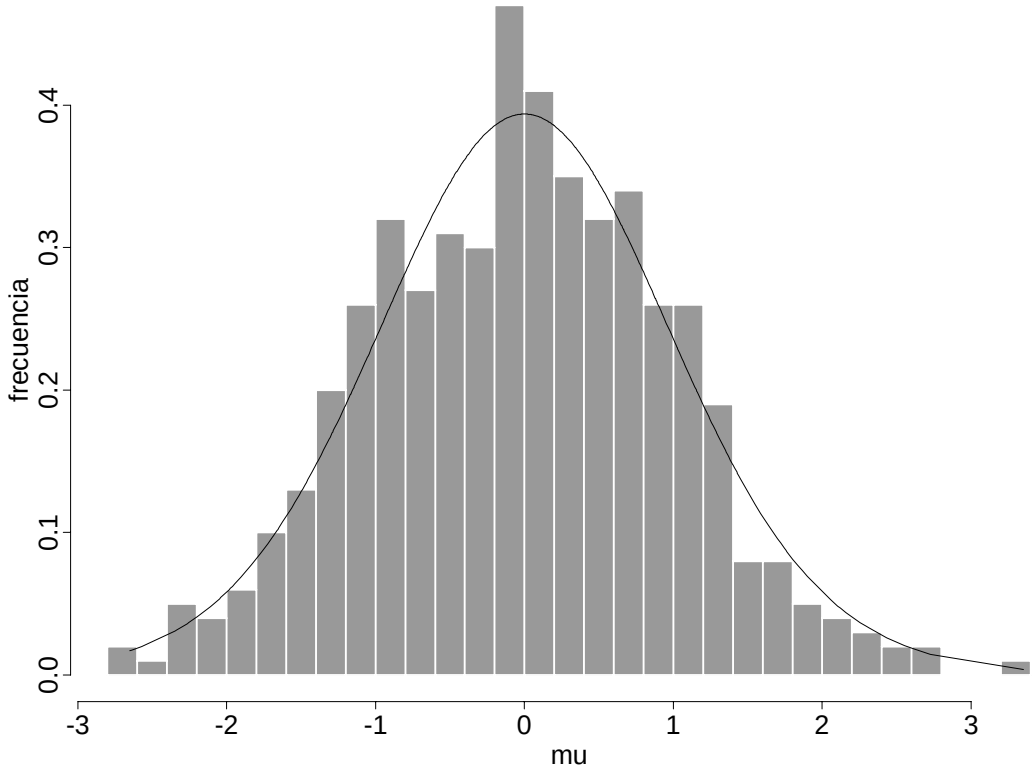
Ejemplo 112 *Sean* $n = 20$, $\bar{x} = 0$ *y* $s^2 = 20$. *Entonces, dada la distribución a priori de Jeffreys,*

$$\begin{aligned} \phi|\mathbf{x} &\sim \mathcal{G}\left(\frac{19}{2}, \frac{19 \times 20}{2}\right) \\ \mu|\mathbf{x}, \phi &\sim \mathcal{N}(0, 1/(20\phi)) \end{aligned}$$

Se hace una muestra Monte-Carlo de tamaño 500 de la distribución a posteriori.

El diagrama ilustra un histograma de los valores de μ *muestreados. Es similar a la densidad* t_{19} *que es la verdadera densidad a posteriori de* μ.

Observación 72 *Se mejora la aproximación utilizando muestras Monte-Carlo más grandes.*



Se estima, por ejemplo, un intervalo de credibilidad para μ *utilizando los cuantiles de la muestra. El 5 % cuantil es* $-1{,}67$ *y el 95 % cuantil es 1.63. El verdadero 95 % cuantil de* t_{19} *es* $1{,}73$. *La media muestral es* ,022, *bastante parecida a* 0, *la verdadera media a posteriori.*

La función de importancia

Son poco comunes las situaciones en las que se puede muestrear la distribución a posteriori fácilmente. Es más habitual que sólo se conozca el núcleo de la distribución, faltando la constante de integración. En estas situaciones, se modifica el método de Monte Carlo.

Pongamos una distribución bien conocida $h(\theta)$, **la función de importancia**. Entonces tenemos

$$E[g(\theta|\mathbf{x})] = \frac{\int w(\theta)g(\theta)h(\theta)\ d\theta}{\int w(\theta)h(\theta)\ d\theta}$$

donde el peso $w(\theta)$ es

$$w(\theta) = \frac{f(\theta)l(\theta|x)}{h(\theta)}.$$

Dados unos datos muestreados de $h(\theta)$, se aproxima la esperanza con

$$E[g(\theta|\mathbf{x})] \approx \frac{\sum_{i=1}^{N} w_i g(\theta^{(i)})}{\sum_{i=1}^{N} w_i}$$

donde

$$w_i = \frac{f(\theta^{(i)}) l(\theta^{(i)}|\mathbf{x})}{h(\theta^{(i)})}$$

Observación 73 *En teoría, se puede usar cualquier función de importancia $h(\theta)$ pero si $h(\theta)$ es muy diferente a $f(\theta|\mathbf{x})$ se necesitaran muestras muy grandes (porque se van a tener muchos elementos en la cola de la distribución a posteriori con pesos muy bajos, y unos pocos datos con pesos altos).*

Por tanto el método funciona mejor si $h(\theta)$ está próxima a $f(\theta|\mathbf{x})$, Elecciones posibles son $h(\theta) = f(\theta)$ o (mejor) $h(\theta) \propto l(\theta|\mathbf{x})$,

Si $h(\theta) = f(\theta|\mathbf{x})$ el método se reduce al método Monte-Carlo básico.

Ejemplo 113 *Volviendo al Ejemplo 110 utilizaremos el método Monte-Carlo con muestreo de importancia para estimar la media a posteriori.*

La media y varianza muestral son $\bar{x} = 10{,}56$ *y* $s^2 = 5{,}44$*. Utilizamos como función de importancia la densidad normal:* $h(\theta) = \mathcal{N}(\bar{x}, s^2/2) = \mathcal{N}(5{,}78, 2{,}72)$.

Observación 74 *A menudo se utilizan funciones de importancia con colas algo más anchas que la verdadera distribución real.*

Entonces se tiene $w(\theta) \propto \frac{\exp\left\{\frac{1}{5{,}44}(\theta - 10{,}56)^2\right\}}{\prod_{i=1}^{5}(1+(\theta - x_i)^2)}$.

La tabla ilustra los estimadores para muestras Monte-Carlo de varios tamaños.

N	$E[\theta\|\mathbf{x}]$
10	10,80
100	10,70
1000	10,60
10000	10,62
100000	10,62

(Se hicieron 100000 iteraciones en aproximadamente 5 segundos en unas estaciones de trabajo antiguas)

La media estimada es igual a la media estimada mediante la regla de Simpson.

El método de rechazo

La idea es muestrear una distribución $h(\theta)$ (una función envolvente) con la propiedad de que

$$f(\theta)l(\theta|\mathbf{x}) < Mh(\theta)$$

para una constante $M > 0$. El algoritmo es

1. $i = 1$,

2. Generar $\theta^{(i)}$ de $h(\cdot)$,

3. Generar $U \sim \mathcal{U}(0, 1)$,

4. Si $U < \frac{f(\theta^{(i)})l(\theta^{(i)}|\mathbf{x})}{Mh(\theta^{(i)})}$ aceptar $\theta^{(i)}$ y si no, rechazarlo,

5. Continuar hasta que la muestra sea completa.

Ejemplo 114 *Consideramos el problema de generar una muestra de una distribución normal* $\mathcal{N}(0,1)$ *truncado para que* $\theta > 0$.

$$f(\theta) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-\theta^2/2} & \text{si } \theta > 0, \\ 0 & \text{si no} \end{cases}$$

Usamos cómo función envolvente una densidad exponencial:

$$h(\theta) = e^{-\theta} \qquad \text{si } \theta > 0.$$

Se necesita encontrar el valor de M *para que*

$$\begin{aligned} \frac{2}{\sqrt{2\pi}} e^{-\theta^2/2} &< M e^{-\theta} \quad \forall\, \theta > 0 \\ M &> \sqrt{\frac{2}{\pi}} e^{\theta - \theta^2/2} \\ M &> \sqrt{\frac{2}{\pi}} e^{1/2} \end{aligned}$$

tomando el máximo cuando $\theta = 1$, *que implica* $M > 1{,}3155$.

Elegimos $M = 1{,}315$ *(el más eficiente) y entonces se aceptan los valores de* θ *generados con probabilidad*

$$p = e^{-\theta^2/2+\theta-1/2} \leq 1.$$

El siguiente diagrama muestra la verdadera densidad semi normal y la densidad envolvente.

Se generó una muestra de 300 datos de la distribución exponencial y se rechazaron 79 datos. El diagrama muestra un histograma de los datos aceptados y la verdadera densidad normal. El ajuste es bastante razonable.

Observación 75 *Propiedades y problemas son*

- $h(\cdot)$ *debe ser parecida a* $f(\theta|\mathbf{x})$*, pero con colas mas largas,*
- *Funciona mejor cuando* M *es pequeño,*
- *Si* $h(\theta) = f(\theta|\mathbf{x})$ *tenemos el método Monte-Carlo simple,*
- *A menudo es difícil elegir* $h(\cdot)$ *y* M*.*

Otros métodos

- Bootstrap. Ver por ejemplo Rubin (1981).

- El algoritmo EM para estimar la moda de la distribución a posteriori. (Dempster et al 1977).

- Métodos MCMC. Ver el capítulo 10.

Aplicación a inferencia para el sistema de colas $Er/M/1$

Consideramos inferencia para el sistema de colas $Er/M/1$. Ver Wiper (1997).

Esta cola tiene servicios Markovianos (media $1/\mu$) y tiempos entre llegadas distribuidos como $\mathcal{G}(k, k\lambda)$. Esta distribución se llama la distribución Erlang, con k etapas exponenciales, cada una de media $1/\lambda k$..

Observación 76 *Incluye la cola* $M/M/1$ *(*$k = 1$*) y la cola* $M/D/1$ *con servicios determinísticos (*$k \to \infty$*).*

Inferencia Bayesiana para este sistema

Dada una muestra de tiempos entre llegadas $\mathbf{x}$ y tiempos de servicio $\mathbf{y}$, se puede demostrar que, dadas las distribuciones a priori de Jeffreys para λ, $\mu|k$ y k, las distribuciones a posteriori serán

$$f(k, \lambda, \mu|\mathbf{x}, \mathbf{y}) \propto f(k, \lambda|\mathbf{x}) f(\mu|\mathbf{y})$$

donde $\mu|\mathbf{y} \sim \mathcal{G}(n_s, t_s)$ y $\lambda|k, \mathbf{x} \sim \mathcal{G}(kn_l, kt_l)$ y

$$P(k|\mathbf{x}) \propto \frac{\Gamma(n_l k)}{\Gamma(k)^{n_l}} \left(\frac{T_l}{t_l^{n_l}}\right)^{k-1}.$$

La distribución estacionaria del número de clientes en el sistema

Suponiendo que la intensidad de tráfico $\rho = \lambda/\mu < 1$, existe una distribución estacionaria del tamaño de la cola N.

En este caso, condicionada en los parámetros, la distribución es geométrica $\pi(N|k,\nu) = (1-\nu)\nu^N$, donde

$$\nu\left(1-\frac{\nu-1}{k\rho}\right)^k = 1.$$

Ver, por ejemplo Nelson (1995).

Observación 77 *Se necesitan métodos numéricos para estimar el valor de ν, por ejemplo el método Newton-Raphson.*

Observación 78 *La distribución $\pi(N|k,\nu)$ es la distribución vista por una persona llegando en equilibrio.*

La distribución predictiva

Para calcular la distribución predictiva, se emplea una muestra Monte-Carlo. Dadas las probabilidades $P(k|\mathbf{x})$, un algoritmo para estimar la distribución estacionaria es:

1. Fijar $k = 1$,
2. Generar una muestra $(\lambda_k^{(1)}, \mu_k^{(1)}), \ldots, (\lambda_k^{(S)}, \mu_k^{(S)})$ para S grande.
3. Rechazar y remuestrear si $\rho^{(i)} \geq 1$,
4. Calcular $\nu_k^{(i)}$, $i = 1, \ldots, S$
5. Estimar $\pi(N|k, \mathbf{x}, \mathbf{y})$ mediante $\widehat{\pi}(N|k, \mathbf{x}, \mathbf{y}) = \frac{1}{S} \sum_{i=1}^{S} \pi(N|k, \nu^{(i)})$
6. $k = k + 1$. Ir a 2.

Ahora, se estima la distribución del tamaño de la cola con

$$\pi(N|\mathbf{x}, \mathbf{y}) \approx \sum_k P(k|\mathbf{x}) \widehat{\pi}(N|k, \mathbf{x}, \mathbf{y})$$

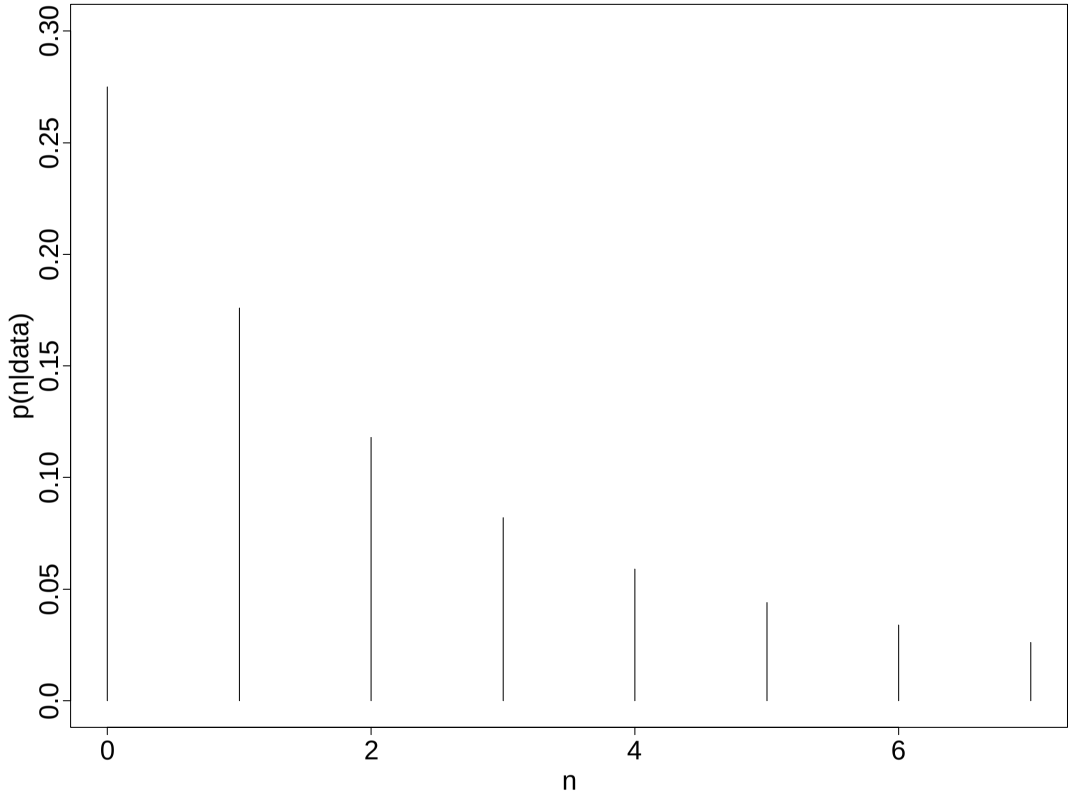
Observación 79 *Se aproximan las probabilidades* $P(k|\mathbf{x})$ *truncando la suma.*

Ejemplo 115 *Se supone que se observan 51 tiempos entre llegadas y 41 tiempos de servicio. Los datos son* $t_s = 16{,}70$, $t_l = 24{,}34$ *y* $\log T_l = -41{,}08$. *(Se han generado estos datos de* $\mathcal{E}(2{,}2)$ *y* $\mathcal{G}(6, 2)$ *).*

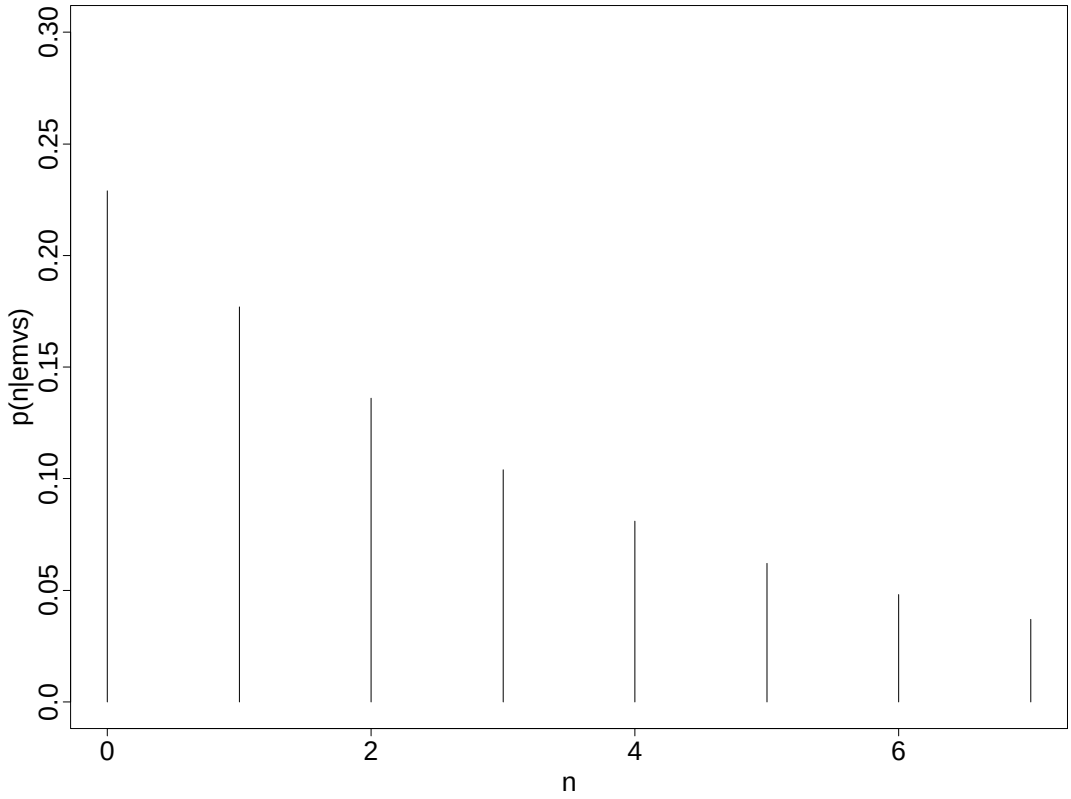
Se supone un soporte de $[1, 20]$ *para* k, *y distribuciones a priori propias pero poco informativas para todos los parámetros.*

Dadas estas distribuciones a priori se tiene $P(\rho < 1) = ,4$ *y* $E[k] = 10{,}3$.

Después de ver los datos se tiene $P(\rho < 1|\mathbf{x}, \mathbf{y}) = ,8$ *y la moda a posteriori de* k *es 5. La figura muestra las probabilidades del tamaño de la cola en equilibrio.*


p(n|data)
0.0
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
n

El siguiente diagrama ilustra las probabilidades estimadas utilizando los EMVs de los parámetros. Se observa que los dos diagramas son muy parecidos.



Comentarios y Extensiones

- Para este sistema la distribución estacionaria del tamaño de la cola visto por un observador externo es distínto.

- También es posible estimar las distribuciones de tiempo de espera o de la duración de un periodo de inactividad.

- Se puede extender el análisis a sistemas con múltiples servidores pero el coste computacional crece rápidamente.