CAPÍTULO 5. ESTIMACIÓN Y CONTRASTES DE HIPÓTESIS

Para leer

Kass y Raftery (1995).

Gelman et al (1995). Capítulo 6, Sección 6.5.

Lee (1997), Capítulo 4.

Berger (1985), Capítulo 4, Secciones 4.3 – 4.4.

En este capítulo se considera el enfoque bayesiano a la estimación puntual y de intervalos y los métodos utilizados para implementar contrastes de hipótesis y selección de modelos.

Estimación puntual

Para los bayesianos, el problema de estimación es un problema de decisión. Asociada con cada estimador T es una pérdida $L(T, \theta)$ que refleja la diferencia entre θ y el estimador.

Algunas funciones de pérdida típicamente consideradas son:

- $L(T, \theta) = (T - \theta)^2$, es la pérdida cuadrática
- $L(T, \theta) = |T - \theta|$ es la pérdida lineal absoluta
- $L(T, \theta) = \begin{cases} 0 & \text{si } T = \theta \\ 1 & \text{si } T \neq \theta \end{cases}$, es la pérdida "todo o nada".

Suponiendo la distribución (a priori) $f(\theta)$ para θ, entonces es posible calcular la pérdida esperada asociada con un estimador T.

Definición 8 *Para un estimador T, el* **riesgo Bayes** *es la pérdida esperada*

$$R(T) = E\left[R(T,\theta)\right] = \int L(T,\theta)f(\theta)\,d\theta.$$

Parece natural elegir el estimador que minimíza el riesgo Bayes. Esta estimador es el estimador Bayes.

Definición 9 *El* **estimador Bayes** *B es una solución de*

$$B = \min_T R(T)$$

Obviamente, después de observar una muestra de datos, se calcularía el estimador Bayes basada en suponer la distribución a posteriori $f(\theta|\mathbf{x})$.

La media, mediana y moda son estimadores Bayes bajo las funciones de pérdida cuadrática, lineal y todo o nada respetivamente.

Ejemplo 63 *Dada la pérdida cuadrática, se tiene*

$$\begin{aligned} E[L(T,\theta)] &= \int (T-\theta)^2 f(\theta)\, d\theta \\ &= \int (T - E[\theta] + E[\theta] - \theta)^2 f(\theta)\, d\theta \\ &= \int \{(T-E[\theta])^2 + (E[\theta]-\theta)^2\} f(\theta)\, d\theta \\ &= V[\theta] + (T - E[\theta])^2 \end{aligned}$$

y $B = E[\theta]$ *es el estimador Bayes.*

Ejemplo 64 *Con la pérdida lineal absoluta,*

$$\begin{aligned}
E[L(T,\theta)] &= \int |T-\theta| f(\theta)\, d\theta \\
&= \int_{-\infty}^{T} (T-\theta) f(\theta)\, d\theta + \\
&\quad + \int_{T}^{\infty} (\theta - T) f(\theta)\, d\theta \\
\frac{d}{dT} E[L(T|\theta)] &= (T-T)f(T) + \int_{-\infty}^{T} f(\theta)\, d\theta - \\
&\quad -(T-T)f(T) - \int_{T}^{\infty} f(\theta)\, d\theta \\
&= F(T) - (1 - F(T)) \\
&= 2F(T) - 1 \\
0 &= 2F(B) - 1 \\
F(B) &= 1/2
\end{aligned}$$

y el estimador Bayes es la mediana de la distribución de θ.

Ejemplo 65 *Siendo θ discreta, dada la pérdida "todo o nada" se tiene*

$$\begin{aligned} E[L(T,\theta)] &= \sum_{\theta \neq T} P(\theta) \\ &= P(\theta \neq T) \end{aligned}$$

y se minimiza la pérdida esperada eligiendo el estimador Bayes B como la moda de la distribución de θ.

Observación 44 *En el caso continuo, dada la pérdida*

$$L(T,\theta) = \begin{cases} 0 & \textit{si } |T-\theta| < \epsilon \\ 1 & \textit{si } |T-\theta| \geq \epsilon \end{cases}$$

se tiene

$$E[L(T,\theta)] = P(|T-\theta| \geq \epsilon) = 1 - P(T-\epsilon < \theta < T+\epsilon)$$

que implica que B es el centro del intervalo más probable (intervalo modal) de anchura 2ϵ. Dejando que $\epsilon \to 0$, entonces $B \to$ la moda.

Discusión sobre estimación puntual

- Parece muy artificial el uso de estimadores puntuales dado que se puede presentar la distribución (a posteriori) entera. Ver Box y Tiao (1973).

- En un problema real, sería casi imposible decidir sobre la forma de la función de pérdida. Esta función debe representar pérdidas económicas reales de alguna manera.

- Las funciones cuadráticas y lineales dan la posibilidad de pérdidas infinitas que no parece muy razonable.

- Ver Robert (2004) para más detalles.

Estimación por intervalos

Hemos visto intervalos de credibilidad anteriormente. Sigue la definición formal.

Definición 10 *Si $f(\theta)$ es la densidad (a priori) para θ, se dice que (a, b) es* **un intervalo de credibilidad** *(a priori) de $100 \times (1 - \alpha)\,\%$ si*

$$P(a \leq \theta \leq b) = 1 - \alpha$$

Ejemplo 66 *$X|\mu \sim \mathcal{N}(\mu, 1)$. Sea $f(\mu) \propto 1$, entonces, $\mu|\mathbf{x} \sim \mathcal{N}(\bar{x}, 1/n)$ y algunos intervalos de credibilidad (a posteriori) de 95 % son*

$$(-\infty, \bar{x} + 1{,}64/\sqrt{n}) \quad \textit{o} \quad (\bar{x} - 1{,}64/\sqrt{n}, \infty) \quad \textit{o} \quad (\bar{x} \pm 1{,}96/\sqrt{n})$$

Existen muchos (infinitos) intervalos de credibilidad posibles. El intervalo más corto se llama **un intervalo de máxima densidad** (a priori) (MD)

Definición 11 *El intervalo MD de* $100 \times (1 - \alpha)$ % *es el intervalo de forma*

$$C = \{\theta : f(\theta) \geq c(\alpha)\}$$

donde $c(\cdot)$ *es la constante más grande para que*

$$P(C) \geq 1 - \alpha$$

Ejemplo 67 *Volviendo al Ejemplo 66, el intervalo MD a posteriori de 95 % es*

$$\bar{x} \pm 1{,}96/\sqrt{n}$$

Observación 45 *Es fácil generalizar el intervalo de credibilidad a situaciones multivariables* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$*. En este caso, se define una región de credibilidad* $\mathbf{C}$ *como*

$$P(\boldsymbol{\theta} \in \mathbf{C}) = 1 - \alpha.$$

Contrastes de hipótesis

Consideramos dos hipótesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ frente $H_1 : \boldsymbol{\theta} \in \Theta_1$, donde $\Theta_0 \cup \Theta_1 = \Theta$ y $\Theta_0 \cap \Theta_1 = \phi$.

En teoría, es fácil distinguir entre las dos hipótesis. Dadas las probabilidades a posteriori $P(H_0|\mathbf{x})$ y $P(H_1|\mathbf{x})$ se tiene el problema de decisión de elegir una de las dos hipótesis.

Suponiendo que se puede decir una función de pérdida para representar las pérdidas asociadas con los errores de tipo I o de tipo II, se elige la hipótesis que minimiza el riesgo Bayes.

Ejemplo 68 *Dada la pérdida "todo o nada",*

$$L(H_i, \theta) = \begin{cases} 0 & \textit{si } H_i \textit{ es verdadero} \\ 1 & \textit{si } H_{1-i} \textit{ es verdadero} \end{cases}$$

para $i = 0, 1$, *se acepta* H_0 *si* $P(H_0|\mathbf{x}) > P(H_1|\mathbf{x})$.

Si se quiere hacerlo más difícil rechazar la hipótesis nula es posible utilizar una función de pérdida asimétrica.

Ejemplo 69 *Siendo la función de pérdida*

$$L(H_0, \theta) = \begin{cases} 0 & \textit{si } H_0 \textit{ es verdadero} \\ \alpha & \textit{si } H_1 \textit{ es verdadero} \end{cases}$$

$$L(H_1, \theta) = \begin{cases} 1-\alpha & \textit{si } H_0 \textit{ es verdadero} \\ 0 & \textit{si } H_1 \textit{ es verdadero} \end{cases}$$

se rechaza H_0 *si* $R(H_0) > R(H_1)$, *es decir si*

$$\begin{aligned} \alpha(1 - p(H_0|\mathbf{x})) &> (1-\alpha)P(H_0|\mathbf{x}) \\ P(H_0|\mathbf{x}) &< \alpha \end{aligned}$$

Ejemplo 70 *Sea* $X|\theta \sim N\left(\theta, \frac{1}{\phi}\right)$. *Se quiere contrastar* $H_0 : \theta \leq 0$ *frente* $H_1 : \theta > 0$. *Dada una distribución a priori uniforme,* $f(\theta) \propto 1$ *se tiene* $\theta|\mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{1}{n\phi}\right)$. *Entonces,*

$$\begin{aligned} P(H_0|\mathbf{x}) &= P(\theta \leq 0|\mathbf{x}) \\ &= P\left(\sqrt{n\phi}(\theta - \bar{x}) \leq -\sqrt{n\phi}\bar{x}|\mathbf{x}\right) \\ &= \Phi\left(-\sqrt{n\phi}\bar{x}\right) \end{aligned}$$

donde $\Phi(\cdot)$ *es la función de distribución normal.*

Esta probabilidad es igual al p valor clásico para el contraste $H_0 : \theta \leq 0$ *frente* $H_1 : \theta > 0$.

$$\begin{aligned} P(\bar{X} \geq \bar{x}|H_0) &= P\left(\sqrt{n\phi}\bar{X} \geq \sqrt{n\phi}\bar{x}|H_0\right) \\ &= 1 - \Phi\left(\sqrt{n\phi}\bar{x}\right) \\ &= \Phi\left(-\sqrt{n\phi}\bar{x}\right) \end{aligned}$$

Ejemplo 71 *Sea $X|\lambda \sim \mathcal{E}(\lambda)$ con las hipótesis $H_0 : \lambda \leq \lambda_0$ y $H_1 : \lambda > \lambda_0$. Suponiendo una distribución a priori impropia $f(\lambda) \propto \frac{1}{\lambda}$, la distribución a posteriori dada una muestra de tamaño n es $\mathcal{G}(n, n\bar{x})$ y luego*

$$P(H_0|\mathbf{x}) = \int_0^{\lambda_0} g(\lambda|n, n\bar{x})\, d\lambda$$

donde $g(\lambda|a, b)$ representa la función de densidad gamma con parámetros a y b.

En el caso clásico, se sabe que $\sum_{i=1}^{n} X_i \sim \mathcal{G}(n, \lambda)$ y luego que bajo H_0, $S = \bar{X} \sim \mathcal{G}(n, n\lambda_0)$. Entonces, el p-valor clásico es

$$\begin{aligned} P(S < \bar{x}|H_0) &= \int_0^{\bar{x}} g(s|n, n\lambda_0)\, ds \\ &= \int_0^{\lambda_0} g(z|n, n\bar{x})\, dz \end{aligned}$$

(definiendo $Z = \lambda_0 S/\bar{x}$) igual a la probabilidad a posteriori bayesiana.

Ejemplo 72 *Volvemos al Ejemplo 4. Se tiene una moeda con $p = P($cruz$)$. Se quiere contrastar $H_0 : p \leq 0{,}5$ frente a $H_1 : p > 0{,}5$. Se lanza la moneda 12 veces observando 9 cruces.*

La primera tabla muestra los p-valores clásicos para los experimentos binomial y binomial negativa.

Experimento	*P-valor*
Binomial	0,0730
Binomial Negativa	0,0327

La siguiente tabla muestra las probabilidades a posteriori de H_0 dadas varias distribuciones a priori.

A priori	$P(H_0\|\mathbf{x})$
Jeffreys Bi $p \sim \mathcal{B}(0{,}5, 0{,}5)$	0,0260
Jeffreys NegBi $f(p) \propto \frac{1}{p\sqrt{1-p}}$	0,0197
Uniforme	0,0461
Haldane $f(p) \propto \frac{1}{p(1-p)}$	0,0327

El resultado suponiendo la distribución a priori de Haldane coincide con el p-valor suponiendo un experimento binomial negativa.

Observación 46 *Es posible demostrar la coincidencia de las probabilidades teóricamente.*

Comentarios

- Formalmente, no se han definido bien las probabilidades a priori de H_0 y H_1 en los ejemplos anteriores. En teoría se debe definir las probabilidades $P(H_0)$ y $P(H_1)$ y densidades a priori $f(\theta|H_0)$ y $f(\theta|H_1)$. Se puede interpretar los resultados de los ejemplos como límites de probabilidades dada una clase de distribuciones a priori.

- Se ha visto en los ejemplos anteriores que si la hipótesis nula y la hipótesis alternativa son compuestas, hay una coincidencia en muchos casos entre el p-valor y la probabilidad a posteriori de H_0 suponiendo alguna distribución a priori impropia. Ver Casella y Berger (1987) para una formalización.

En el caso de que la hipótesis nula sea puntual $H_0 : \theta = \theta_0$ con la alternativa $H_1 : \theta \neq \theta_0$ pueden existir grandes diferencias.

La paradoja de Lindley/Jeffreys

Ejemplo 73 *Sea* $X|\theta \sim \mathcal{N}(\theta, 1)$*.*

Se quiere contrastar $H_0 : \theta = 0$ *frente a la alternativa* $H_1 : \theta \neq 0$*.*

Sean las probabilidades a priori

$$f_0 = P(H_0) = 0{,}5 = P(H_1) = f_1$$

con la distribución $\theta|H_1 \sim \mathcal{N}(0, 1)$*.*

Observación 47 *Es importante que* $f(\theta|H_1)$ *sea propia. Si no*

$$\begin{aligned} P(H_0|\mathbf{x}) &= \frac{P(H_0)f(\mathbf{x}|\theta_0)}{P(H_0)f(\mathbf{x}|\theta_0) + P(H_1)f(\mathbf{x}|H_1)} \\ &= \frac{P(H_0)f(\mathbf{x}|\theta_0)}{P(H_0)f(\mathbf{x}|\theta_0) + P(H_1)\int f(\mathbf{x}|\theta)f(\theta|H_1)\,d\theta} \end{aligned}$$

que no está definida.

Suponiendo que se observa la media $\bar{x}$ de una muestra de tamaño n, entonces

$$l(\theta|\mathbf{x}) = \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2\right)$$

y se pueden evaluar las probabilidades a posteriori, $\alpha_i = P(H_i|\mathbf{x})$.

$$\begin{aligned}
\alpha_0 &= P(H_0|\bar{x}) \\
&= \frac{P(H_0)f(\bar{x}|H_0)}{P(H_0)f(\bar{x}|H_0) + P(H_1)f(\bar{x}|H_1)} \\
&= \frac{f_0 l(\theta = 0|\mathbf{x})}{f_0 l(\theta = 0|\mathbf{x}) + f_1 \int l(\bar{x}|\theta)f(\theta|H_1)\,d\theta} \\
&= \frac{l(\theta = 0|\mathbf{x})}{l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}, H_1)f(\theta|H_1)\,d\theta} \\
&= K\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right)
\end{aligned}$$

para una constante

$$K^{-1} = l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}, H_1)f(\theta|H_1)\,d\theta.$$

Además:

$$\begin{aligned}
l(\theta|\bar{x}, H_1)f(\theta|H_1) &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \times \\
&\quad \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta^2\right) \\
&= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[n(\bar{x}-\theta)^2+\theta^2\right]\right) \\
&= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[(n+1)\left(\theta-\frac{n\bar{x}}{n+1}\right)^2\right.\right. \\
&\quad \left.\left. +\frac{n\bar{x}^2}{n+1}\right]\right)
\end{aligned}$$

$$\int l(\theta|\bar{x}, H_1)f(\theta|H_1)d\theta = \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right)$$

Entonces, se tiene

$$\begin{aligned}
K^{-1} &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right) + \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right) \\
&= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right)\left(1+\frac{1}{\sqrt{n+1}}\exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right) \\
\alpha_0 &= \left\{1+\frac{1}{\sqrt{n+1}}\exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right\}^{-1}
\end{aligned}$$

Sea $\bar{x} = 2/\sqrt{n} > 1{,}96/\sqrt{n}$.

Para el contraste clásico de H_0 *frente* H_1 *con un nivel de significación de 95 %, el resultado es significativo, y se rechaza la hipótesis* H_0.

Pero en este caso, la probabilidad a posteriori es

$$\begin{aligned} \alpha_0 &= \left\{1 + \frac{1}{\sqrt{n+1}} \exp\left(\frac{n}{n+1}\right)\right\}^{-1} \\ &\to 1 \quad \textit{cuando } n \to \infty. \end{aligned}$$

Una muestra que nos llega a rechazar H_0 *con un contraste clásico nos proporciona una probabilidad a posteriori de* H_0 *que se acerca a* 1. *Esta paradoja se llama la paradoja de Lindley y Jeffreys. Ver Lindley (1957) o Jeffreys (1961).*

Observación 48 *Se puede pensar que la elección de la distribución a priori de $\theta|H_1$ es muy influyente. No obstante para cualquiera distribución a priori propia $f(\theta|H_1)$, se tiene*

$$f(\bar{x}|H_1) = \int f(\theta|H_1) f(\bar{x}|\theta)\, d\theta \leq f(\bar{x}|\widehat{\theta})$$

donde $\widehat{\theta} = \bar{x}$ es el EMV. Luego

$$\begin{aligned} \frac{\alpha_0}{\alpha_1} &= \frac{P(H_0|\bar{x})}{P(H_1|\bar{x})} \\ &\geq \frac{P(H_0) f(\bar{x}|\theta = 0)}{P(H_1) f(\bar{x}|\widehat{\theta})} \\ &= \exp\left(-\frac{n\bar{x}^2}{2}\right) \end{aligned}$$

y $\alpha_0 \geq \frac{\exp\left(-\frac{n\bar{x}^2}{2}\right)}{1+\exp\left(-\frac{n\bar{x}^2}{2}\right)} = \left\{1 + \exp\left(\frac{n\bar{x}^2}{2}\right)\right\}^{-1}.$

Siendo $\bar{x} = \frac{2}{\sqrt{n}}$, para cualquier n, la probabilidad a posteriori de H_0 es $P(H_0|\bar{x}) > 0{,}1192$. El p-valor sobre-estima la evidencia contra H_0.

Factores Bayes

Ejemplo 74 *Un problema importante es la elección de un modelo M_i de una clase $\mathcal{M} = \{M_i : i = 1, \ldots, k\}$. En este caso*

$$P(M_i|\mathbf{x}) = \frac{P(M_i) f(\mathbf{x}|M_i)}{\sum_{j=1}^{k} P(M_j) f(\mathbf{x}|M_j)},$$

y bajo la pérdida "todo o nada", se elige el modelo más probable.

Pero las probabilidades a priori son muy influyentes. Si la clase $\mathcal{M}$ de modelos es grande, a menudo es difícil especificar precisamente las probabilidades $P(M_i)$.

Entonces, se introduce otro concepto con menos dependencia de la distribución a priori. Se consideran dos hipótesis (o modelos) H_0 y H_1.

Sean $f_0 = P(H_0)$ y $f_1 = P(H_1)$ y $\alpha_0 = P(H_0|\mathbf{x})$ y $\alpha_1 = P(H_1|\mathbf{x})$.

Definición 12 *El* **factor Bayes** *a favor de H_0 es*

$$B = \frac{\alpha_0/\alpha_1}{f_0/f_1} = \frac{\alpha_0 f_1}{\alpha_1 f_0}$$

Ver (Jeffreys 1961).

Observación 49 *El factor Bayes representa las posibilidades (*odds*) a posteriori divididos por los posibilidades a priori. Nos informe de los cambios en nuestras creencias causados por los datos.*

Es casi objetiva y parcialmente elimina la influencia de la distribución inicial.

Observación 50 *La definición extiende al caso de más de dos hipótesis. En la situación del Ejemplo 74 se define*

$$B_j^i = \frac{P(M_i|\mathbf{x})P(M_j)}{P(M_j|\mathbf{x})P(M_i)}$$

es al factor Bayes a favor del modelo M_i frente al modelo M_j.

Ejemplo 75 *Consideramos el caso de dos hipótesis simples.*

Sea $H_0 : \theta = \theta_0$ *frente* $H_1 : \theta = \theta_1$. *Entonces*

$$\begin{aligned}\alpha_0 = P(H_0|\mathbf{x}) &= \frac{f_0 l(\theta_0|\mathbf{x})}{f_0 l(\theta_0|\mathbf{x}) + f_1 l(\theta_1|\mathbf{x})} \\ \alpha_1 = P(H_1|\mathbf{x}) &= \frac{f_1 l(\theta_1|\mathbf{x})}{f_0 l(\theta_0|\mathbf{x}) + f_1 l(\theta_1|\mathbf{x})}\end{aligned}$$

El factor Bayes es

$$\begin{aligned}B &= \frac{\alpha_0}{\alpha_1}\frac{f_1}{f_0} \\ &= \frac{f_0 l(\mathbf{x}|\theta_0)}{f_1 l(\mathbf{x}|\theta_1)}\frac{f_1}{f_0} \\ &= \frac{l(\theta_0|\mathbf{x})}{l(\theta_1|\mathbf{x})}\end{aligned}$$

que coincide con la razón de verosimilutudes. Entonces, la distribución a priori no influye en el factor Bayes.

Ejemplo 76 *Se observa un dato de una distribución exponencial con densidad*

$$f(x|\lambda) = \lambda e^{-\lambda x}.$$

Se quiere contrastar $H_0 : \lambda = 6$ frente a la alternativa $H_1 : \lambda = 3$. ¿Qué es el factor Bayes?

$$\begin{aligned} B &= \frac{l(\lambda = 6|x)}{l(\lambda = 3|x)} \\ &= \frac{6e^{-6x}}{3e^{-3x}} \\ &= 2e^{-3x} \end{aligned}$$

Para cualquier x, se tiene $0 < B < 2$.

Para contrastes compuestos, el factor Bayes depende de las distribuciones a priori.

Ejemplo 77 *Volviendo al Ejemplo 74, sea*

$$X|M_i \sim f(\cdot|\boldsymbol{\theta}_i) \text{ para } i = 1, 2, \ldots$$

con distribuciones a priori $\boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}_i|M_i)$. *El factor Bayes a favor de* M_i *frente* M_j *es*

$$\begin{aligned} B_j^i &= \frac{P(M_i|\mathbf{x})P(M_j)}{P(M_j|\mathbf{x})P(M_i)} \\ &= \frac{\frac{f(\mathbf{x}|M_i)P(M_i)}{\sum_{l=1}^k f(\mathbf{x}|M_l)P(M_l)} P(M_j)}{\frac{f(\mathbf{x}|M_j)P(M_j)}{\sum_{l=1}^k f(\mathbf{x}|M_l)P(M_l)} P(M_i)} \\ &= \frac{f(\mathbf{x}|M_i)}{f(\mathbf{x}|M_j)} \\ &= \frac{\int f(\mathbf{x}|\boldsymbol{\theta}_i)f(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i}{\int f(\mathbf{x}|\boldsymbol{\theta}_j)f(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j} \end{aligned}$$

que depende de las densidades a priori $f(\boldsymbol{\theta}_i|M_i)$ *y* $f(\boldsymbol{\theta}_j|M_j)$ *elegidas.*

Convergencia del factor Bayes

El factor Bayes es consistente. Si H_0 es verdadero, entonces $B \to \infty$ cuando $n \to \infty$ y si H_1 es verdadero, $B \to 0$ cuando $n \to \infty$.

Ejemplo 78 *Volviendo al Ejemplo 76, supongamos ahora que se observa una muestra* $\mathbf{x}$ *de tamaño* n*. Entonces el factor Bayes a favor de* H_0 *es*

$$B = 2^n \exp\left(-3n\bar{x}\right).$$

Si H_0 *es verdadero entonces,*

$$\bar{x} \to \frac{1}{6} \text{ cuando } n \to \infty, \text{ es decir}$$

$$B \to 2^n e^{-n/2} = \exp\left(n\left(\log 2 - \frac{1}{2}\right)\right) \to \infty$$

Escalas de Evidencia

Se puede interpretar el estadístico $2 \log B$ como una versión bayesiana del estadístico clásico del logaritmo de verosimilitudes. Ver Kass y Raftery (1995). Se sugiere la siguiente tabla que refleja la evidencia en contra de H_1.

$2 \log_{10} B$	B	Evidencia contra H_1.
0 a 2	1 a 3	Sólo merece comentar
2 a 6	3 a 20	Positiva
6 a 10	20 a 150	Fuerte
> 10	> 150	Muy fuerte

(Jeffreys (1961) propone otra tabla de evidencia parecida.)

Ejemplo 79 *En el Ejemplo 76 sea $x = 0{,}5$. Entonces*

$$\begin{aligned} B &= 2e^{-1,5} \\ 2 \log_{10} B &\approx -0{,}7 \\ &\Rightarrow \textit{muy poca evidencia contra } H_0 \end{aligned}$$

Cálculo del factor Bayes y su relación con otros criterios

Volvemos al problema de la selección de un modelo del Ejemplo 77. A menudo, es muy difícil calcular las distribuciones marginales

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

necesarios para la computación del factor Bayes, especialmente si la dimensión del espacio paramétrico es alto.

Existen alternativas más faciles de evaluar.

- El critero Schwarz (Schwarz 1978) para evaluar la bondad de ajuste de un modelo M_i es

  $$S_i = -\log f(\mathbf{x}|\widehat{\boldsymbol{\theta}}_i, M_i) + \frac{1}{2} d_i \log n$$

  donde $\widehat{\boldsymbol{\theta}}_i$ es el EMV bajo M_i y d_i es la dimensión de $\boldsymbol{\theta}_i$.

Se puede demostrar que cuando $n \to \infty$ se tiene

$$\frac{S_i - S_j - \log B_j^i}{\log B_j^i} \to 0.$$

Observación 51 *$S_i - S_j$ es una aproximación $O\left(n^{-1/2}\right)$ a* $\log B_j^i$*. Ver Kass y Wasserman (1995). $-2S$ es el BIC o* Bayesian information criterion*.*

El inconveniente del criterio Schwarz es que se tiene que calcular el EMV que no es muy natural dentro de un análisis bayesiano. Otro criterio que proporciona una mejor aproximación al factor Bayes es el DIC o *deviance information criterion*.

- EL DIC

Para un modelo con parámetros $\boldsymbol{\theta}$, se define la desviación (deviance) como la distribución a posteriori del logaritmo de la verosimilitud.

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{x}|\boldsymbol{\theta}).$$

Una medida de la bondad de ajuste del modelo es

$$\bar{D} = E[D|\mathbf{x}] = \int D(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}$$

y una medida de la complejidad del modelo es

$$p_D = \bar{D} - D(\bar{\theta})$$

donde $\bar{\boldsymbol{\theta}}$ es un estimador apropiado de $\boldsymbol{\theta}$, por ejemplo la media a posteriori.

Luego se define el DIC como

$$DIC = p_D + \bar{D}.$$

Para comparar dos modelos, simplemente se calcula la diferencia en los DIC.

La ventaja del uso del DIC es que es muy fácil de estimar mediante, por ejemplo una muestra Gibbs. El DIC está implementado en el software Winbugs. Ver Spiegelhalter et al (2002) para más detalles.

Problemas y Generalizaciones del Factor Bayes

Si se usan distribuciones a priori impropias para los parámetros, el factor Bayes no exista.

Ejemplo 80 *Retomando el Ejemplo 77, supongamos que las distribuciones* $f(\theta_i)$ *y* $f(\theta_j)$ *son impropias. Entonces*

$$B = \frac{\int f(\mathbf{x}|\theta_i) f(\theta_i)\, d\theta_i}{\int f(\mathbf{x}|\theta_j) f(\theta_j)\, d\theta_j}$$

no está definido.

En casos así, se generaliza el concepto del factor Bayes. Hay varias posibilidades.

Factores Bayes intrínsecos

Ver: Berger y Perrichi (1996)

Se comparten los datos en $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ donde $\mathbf{y}$ es una muestra de entrenimiento (*training data*).

Se define el factor Bayes parcial a favor del modelo M_i basado en los datos $\mathbf{z}$, después de ver $\mathbf{y}$ como

$$B_j^i(\mathbf{z}|\mathbf{y}) = \frac{\int f(\boldsymbol{\theta}_i|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int f(\boldsymbol{\theta}_j|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_j}$$

que existe si las distribuciones $f(\boldsymbol{\theta}_i|\mathbf{y})$ y $f(\boldsymbol{\theta}_j|\mathbf{y})$ son propias a pesar de que las a priori sean impropias.

Observación 52 *No resuelve todos los problemas porque la partición* $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ *es arbitraria.*

Una posibilidad es definir una media sobre todos los posibles conjuntos $\mathbf{y}$ de mínima dimensión que proporcionan un factor Bayes propio.

Se define

$$B_j^i(G) = \left\{ \prod_{l=1}^{L} B_j^i(\mathbf{z}(l)|\mathbf{y}(l)) \right\}^{1/L} \qquad \text{ó}$$

$$B_j^i(A) = \frac{1}{L} \sum_{l=1}^{L} B_j^i(\mathbf{z}(l)|\mathbf{y}(l))$$

donde el índice l corre por todos los conjuntos $\mathbf{y}(l)$ de mínima dimensión.

Observación 53 *La segunda definición no cumple* $B_j^i(A) = 1/B_j^i(A)$*. Para utilizarlo se necesita que los modelos estén ordenados.*

Ejemplo 81 *Volvemos al Ejemplo 70. Queremos calcular el factor Bayes entre H_0 y H_1.*

Aquí se supusó una distribución a priori uniforme para θ, es decir

$$f(\theta|H_0) \propto 1 \ \ y \ \ f(\theta|H_1) \propto 1.$$

Obviamente el factor Bayes no existe pero si observamos un dato x_l, se tiene

$$\theta|x_l \sim \mathcal{N}(x_l, 1)$$

que es propia, y

$$\begin{aligned} f(\theta|H_0, x_l) &= \frac{f(\theta|x_l)}{P(\theta \leq 0|x_l)} \\ &= \frac{f(\theta|x_l)}{\Phi(-x_l)} \\ f(\theta|H_1, x_l) &= \frac{f(\theta|x_l)}{1 - \Phi(-x_l)} \end{aligned}$$

Entonces, definiendo

$$\mathbf{z}(l) = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_n)$$

se puede calcular el factor Bayes intrínseca.

$$
\begin{aligned}
B(\mathbf{z}(l)|x_l) &= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \times \\
&\quad \frac{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right)\, d\theta}{\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta-x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right)\, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right)\, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right)\, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)}
\end{aligned}
$$

Entonces, el factor Bayes intrínseca es

$$
B_1^0(G) = \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\prod_{l=1}^{n} \frac{1-\Phi(-x_l)}{\Phi(-x_l)}\right)^{\frac{1}{n}}.
$$

Factores Bayes fraccionales

Ver: O'Hagan (1995).

Se define $B_j^i(b) = \frac{g_i(\mathbf{x}|b)}{g_j(\mathbf{x}|b)}$ donde

$$g_i(\mathbf{x}|b) = \int \frac{f(\boldsymbol{\theta}_i) f(\mathbf{x}|\boldsymbol{\theta}_i)}{\int f(\boldsymbol{\theta}_i) \left\{f(\mathbf{x}|\boldsymbol{\theta}_i)\right\}^b \, d\boldsymbol{\theta}_i} \, d\boldsymbol{\theta}_i$$

Observación 54 *El valor de b representa la fracción de los datos elegida para la muestra de entrenimiento. Se puede elegir el mínimo valor posible pero valores más grandes pueden proporcionar resultados más robustos.*

Ejemplo 82 *En el Ejemplo 70, sea $b = 1/n$. Entonces*

$$g_0(\mathbf{x}|b) = \frac{\int_{-\infty}^{0} \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2n}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}$$

$$= \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \Phi\left(-\sqrt{n}\bar{x}\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \Phi\left(-\bar{x}\right)}$$

y parecidamente,

$$g_1(\mathbf{x}|b) = \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \left(1-\Phi\left(-\sqrt{n}\bar{x}\right)\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \left(1-\Phi\left(-\bar{x}\right)\right)}$$

Entonces,

$$B^{1/n} = \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\frac{1-\Phi\left(-\bar{x}\right)}{\Phi\left(-\bar{x}\right)}\right)$$

Observación 55 *Factores Bayes fraccionales y intrínsicas no cumplen con todas las propiedades de factores Bayes simples. Ver O'Hagan (1997).*

Existen varias generalizaciones más del factor Bayes para problemas de selección de modelos. Ver, por ejemplo Perrichi (2005).