Prioris Impróprias e Não Informativas

Prioris Impróprias

No paradigma bayesiano temos que

$$p(\theta \mid x) = \frac{p(\theta)\, p(x \mid \theta)}{\int p(\theta)\, p(x \mid \theta)\, d\theta} \propto p(\theta)\, p(x \mid \theta)\,. \qquad (19)$$

Quando tanto $p(\theta)$ quanto $p(x \mid \theta)$ são densidades, resultados do cálculo de probabilidades mostram que a distribuição a posteriori está bem definida, no sentido que a integral no denominador acima é finita. Veja que nada muda se usarmos acima uma "priori" cuja integral é finita mas não necessariamente igual a um, pois se $\int p(\theta)\, d\theta = I \neq 1$, podemos definir a priori $p^*(\theta) = p(\theta)/I$ e a constante I cancelará no numerador e no denominador (19).

Podemos usar uma priori imprópria?

A situação é diferente se, formalmente, usarmos em (19) uma "densidade" (dita *imprópria* na literatura) cuja integral é infinita, pois nesse caso não é garantido que $\int p(\theta)\, p(x \mid \theta)\, d\theta < \infty$. Esse tipo de "prioris" podem ser pensadas como aproximações para prioris próprias, mas o argumento anterior implica que a propriedade da distribuição a posteriori deve ser analisada caso a caso.

Exemplo 3. Vimos na seção 1.1 o uso da priori "não informativa" $p(\mu) \propto 1$, que é imprópria. A partir da equação (19) temos que,, nesse caso, $p(\mu \mid x_1, \ldots, x_n) \propto \exp\{\frac{-n}{2\sigma^2}(\bar{x} - \mu)^2\}$. Como $\int_{-\infty}^{\infty} \exp\{\frac{-n}{2\sigma^2}(\bar{x} - \mu)^2\}\, d\mu < \infty$, a posteriori resultante será própria qualquer seja a amostra $(x_1, \ldots, x_n)$ observada. Pode-se verificar que o mesmo vale na seção 1.2 para a priori imprópria $p(\mu, \phi) \propto \phi^{-1}$.

Priori imprópria para o Poisson-Gama

Exemplo 4. Suponha que o modelo especifica $x_1, \ldots, x_n \,|\, \theta \sim \text{Poisson}(\theta)$ e queremos usar uma distribuição a priori com média θ_0 e variância tão grande quanto possível. Usando a priori conjugada Gama(α,β) para esse problema, devemos fazer $\alpha = \beta\,\theta_0$ para a restrição da média, em cujo caso a variância será θ_0/β, que é maximizada quando $\beta \downarrow 0$. Na prática, isso significa que a priori deveria ser uma "Gama(0,0)" com densidade proporcional a θ^{-1}. Substituindo em (19) obtemos $p(\theta \,|\, x_1, \ldots, x_n) \propto e^{-n\theta}\theta^{s-1}$. Logo, quando $s = \sum_{i=1}^{n} x_i \geq 1$, a posteriori será uma Gama(s,n), que é própria. Porém, se observarmos $s = 0$, temos que $\int_0^\infty \theta^{-1}\, e^{-n\theta}\, d\theta = \infty$ e a posteriori não vai estar bem definida.

Como dito acima, precisamos avaliar a posteriori caso a caso quando nossa priori é imprópria. Não tem uma regra geral.

O problema de locação

Um modelo de locação específica a observação de uma quantidade X tal que $p(x \mid \theta) = f(x - \theta)$, onde f é uma densidade (própria!) definida em toda a reta real. Esse modelo tem uma simetria fundamental no sentido que é invariante a translações. Mais precisamente, se a é um real fixo e definirmos $Y = X - a$ e $\lambda = \theta - a$, segue do fato que $p_Y(y \mid \theta) = p_X(x = y + a \mid \theta) \, |dx/dy| = f(y + a - \theta) = f(y - \lambda)$ que o modelo $Y \mid \lambda$ também é um modelo de locação. Portanto, se existe nesse caso uma "distribuição a priori não informativa", ela deveria tomar a mesma forma para os modelos $X \mid \theta$ e $Y \mid \lambda$. Em outras palavras, deveríamos ter que $p_\lambda(\lambda) = p_\theta(\lambda)$. Por outro lado, sabemos do cálculo de probabilidades e o fato que $\lambda = \theta - a$ que $p_\lambda(\lambda) = p_\theta(\lambda + a) \, |d\theta/d\lambda| = p_\theta(\lambda + a)$. Juntando os dois resultados, segue que $p_\theta(\lambda) = p_\theta(\lambda + a)$ deveria valer para todo λ e todo a, o que somente é possível se a densidade p_θ for constante (usualmente denotamos $p(\theta) \propto 1$).

O problema de escala

Um modelo de escala específica que $p(x \mid \sigma) = \sigma^{-1} f(x/\sigma)$, onde o parâmetro $\sigma > 0$ e a densidade f é considerada fixa. Esse modelo também tem uma simetria fundamental, no sentido que, se definirmos $Y = X/a$ e $\phi = \sigma/a$ $(a > 0)$, segue que $p_Y(y \mid \sigma) = p_X(ay \mid \sigma) \, |dx/dy| = \phi^{-1} f(y/\phi)$. Dessa forma, tanto o modelo $X \mid \sigma$ quanto o $Y \mid \phi$ são modelos de escala. Se repetimos o argumento que fizemos acima para o modelo de locação, chegaremos ao final que uma distribuição a priori não informativa para o problema de escala deveria satisfazer a condição $p(\sigma) = a^{-1} p(\sigma/a)$ para todo a e σ positivos. É fácil ver que essa condição somente é possível se tivermos que $p(\sigma) \propto \sigma^{-1}$.

(Uma outra forma de obter esse resultado é a seguinte: No modelo de escala com $X > 0$ defina $X^* = \log X$ e $\mu = \log \sigma$ e verifique que o modelo resultante é de locação, de forma que pelo argumento anterior deveríamos ter que $p(\mu) \propto 1$, o que é equivalente a assumir $p(\sigma) \propto \sigma^{-1}$.)

Distribuições a priori de Jeffreys (unidimensional)

Vimos na unidade anterior que um problema fundamental com o Princípio da Probabilidade Inversa de Laplace era o fato de ele não ser invariante a reparametrizações. A partir de essa idéia o físico britânico sugeriu usar como distribuição a priori não informativa uma densidade baseada informação de Fisher. No caso uniparâmetrico, define-se

$$p_J(\theta) \propto I(\theta)^{1/2} = \sqrt{\mathbb{E}_{X\,|\,\theta}\left[\frac{\partial}{\partial\theta}\log p(X\,|\,\theta)\right]^2} = \sqrt{\mathbb{E}_{X\,|\,\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log p(X\,|\,\theta)\right]} \quad (20)$$

Priori de Jeffreys multidimensional

Mais geralmente, quando o espaço paramêtrico tem dimensão r maior do que um, a informação de Fisher é uma matriz quadrada com elementos

$$I_{i,j}(\theta) = \mathbb{E}_{X\,|\,\theta}\left(\frac{\partial}{\partial\theta_i}\log p(X\,|\,\theta)\cdot\frac{\partial}{\partial\theta_j}\log p(X\,|\,\theta)\right) = -\,\mathbb{E}_{X\,|\,\theta}\left(\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log p(X\,|\,\theta)\right)\,.$$

Nesse caso a densidade a priori de Jeffreys é

$$p_J(\theta) \propto \left[\det I(\theta)\right]^{1/2}\,.$$

Reflexões sobre as prioris de Jeffreys

- Se $x_1, \ldots, x_n \,|\, \theta \overset{iid}{\sim} p(x_1 \,|\, \theta)$ e $I_n(\theta)$ é a informação de Fisher associada a essa amostra aleatória de tamanho n, como $I_n(\theta) = n\, I_1(\theta)$, não faz diferença se calcularmos a priori de Jeffreys para toda a amostra ou para uma única observação (i.é. $p_J(\theta) \propto I_n(\theta) \propto I_1(\theta)$).

- Se considerarmos uma reparametrização $\eta = \eta(\theta)$, como $I_\eta(\eta) = I_\theta[\theta(\eta)]\,(d\theta/d\eta)^2$, segue que $p_{J,\eta}(\eta) = p_{J,\theta}[\theta(\eta)]\,|d\theta/d\eta|$, que é a regra do Jacobiano usual e portanto a priori de Jeffreys é invariante sob reparametrizações.

- Alguns autores afirmam que o uso da priori de Jeffreys pode levar a situações anómalas no caso multiparamétrico e sugerem nesse caso usar prioris de Jeffreys calculadas para cada parâmetro por separado e assumir independência. Por exemplo, se $\theta = (\theta_1, \theta_2)$, a sugestão é calcular $p_{J,\theta_i}(\theta_i)$ assumindo o outro parâmetro fixo e depois usar a priori $p_{J,\theta_1}(\theta_1)\, p_{J,\theta_2}(\theta_2)$.

Mais reflexões

- No caso dos modelos de locação e de escala, a priori de Jeffreys coincide com as achadas acima (veja o exemplo 5 abaixo).

- Como o valor esperado na equação 20 é calculado com respeito à distribuição de $x\,|\,\theta$, o uso de prioris de Jeffreys viola o Principio de Verossimilhança.

Exemplo 5. Suponha o modelo de locação $x\,|\,\theta \sim f(x-\theta)$. A Informação de Fisher é

$$I_1(\theta) = \int_{-\infty}^{\infty} \left[\frac{d}{d\theta} \log f(x-\theta)\right]^2 f(x-\theta)\,dx$$
$$= \int_{-\infty}^{\infty} \left[\frac{f'(x-\theta)}{f(x-\theta)}\right]^2 f(x-\theta)\,dx = \int_{-\infty}^{\infty} \left[\frac{f'(u)}{f(u)}\right]^2 f(u)\,du \propto 1\,.$$

Portanto, nesse caso $p_J(\theta) \propto 1$. Um argumento semelhante no modelo de escala mostra que, nesse caso, $p_J(\sigma) \propto \sigma^{-1}$ (faça como exercício).

Dois exemplos

Exemplo 6. Suponha que $x_1, \ldots, x_n \,|\, \theta \sim$ Binomial(n,θ). Como nesse caso temos que $I_1(\theta) = [\theta\,(1-\theta)]^{-1}$, segue que a priori de Jeffreys é $p_J(\theta) \propto [\theta\,(1-\theta)]^{-1/2}$, isto é, a distribuição Beta($\alpha = 1/2$,$\beta = 1/2$).

Exemplo 7. Suponha que $x_1, \ldots, x_n \,|\, \mu, \phi$ Normal(μ,ϕ^{-1}). Como nesse caso temos que $I_1(\mu, \phi) =$, segue que a priori de Jeffreys é $p_J(\mu, \phi) \propto$.