

INFERÊNCIA BAYESIANA

RICARDO S. EHLERS

Primeira publicaç ao em 2002
Segunda edição publicada em 2004
Terceira edição publicada em 2005
Quarta edição publicada em 2006
Quinta edição publicada em 2007

© RICARDO SANDES EHLERS 2003-2011

Sumário

1	Introdução	1
1.1	Teorema de Bayes	1
1.2	Princípio da Verossimilhança	11
1.3	Exercícios	12
2	Distribuições a Priori	14
2.1	Prioris Conjugadas	14
2.2	Conjugação na Família Exponencial	15
2.3	Principais Famílias Conjugadas	19
2.3.1	Distribuição normal com variância conhecida	19
2.3.2	Distribuição de Poisson	20
2.3.3	Distribuição multinomial	21
2.3.4	Distribuição normal com média conhecida e variância desconhecida	22
2.3.5	Distribuição normal com média e variância desconhecidos .	23
2.4	Priori não Informativa	25
2.5	Prioris Hierárquicas	28
2.6	Problemas	30
3	Estimação	35
3.1	Introdução à Teoria da Decisão	35
3.2	Estimadores de Bayes	36
3.3	Estimação por Intervalos	38
3.4	Estimação no Modelo Normal	39
3.4.1	Variância Conhecida	40
3.4.2	Média e Variância desconhecidas	41
3.4.3	O Caso de duas Amostras	42
3.4.4	Variâncias desiguais	45
3.5	Exercícios	47

4 Métodos Aproximados	48
4.1 Computação Bayesiana	48
4.2 Uma Palavra de Cautela	48
4.3 O Problema Geral da Inferência Bayesiana	49
4.4 Método de Monte Carlo Simples	50
4.4.1 Monte Carlo via Função de Importância	54
4.5 Métodos de Reamostragem	57
4.5.1 Método de Rejeição	57
4.5.2 Reamostragem Ponderada	60
4.6 Monte Carlo via cadeias de Markov	63
4.6.1 Cadeias de Markov	63
4.6.2 Acurácia Numérica	64
4.6.3 Algoritmo de Metropolis-Hastings	65
4.6.4 Casos Especiais	71
4.6.5 Amostrador de Gibbs	72
4.7 Problemas de Dimensão Variável	78
4.7.1 MCMC com Saltos Reversíveis (RJMCMC)	81
4.8 Tópicos Relacionados	86
4.8.1 Autocorrelação Amostral	86
4.8.2 Monitorando a Convergência	86
5 Modelos Lineares	88
5.1 Análise de Variância com 1 Fator de Classificação	91
A Lista de Distribuições	93
A.1 Distribuição Normal	93
A.2 Distribuição Log-Normal	94
A.3 A Função Gama	94
A.4 Distribuição Gama	94
A.5 Distribuição Wishart	95
A.6 Distribuição Gama Inversa	95
A.7 Distribuição Wishart Invertida	95
A.8 Distribuição Beta	96
A.9 Distribuição de Dirichlet	96
A.10 Distribuição t de Student	96
A.11 Distribuição F de Fisher	97
A.12 Distribuição de Pareto	97
A.13 Distribuição Binomial	97
A.14 Distribuição Multinomial	97
A.15 Distribuição de Poisson	98
A.16 Distribuição Binomial Negativa	98

<i>SUMÁRIO</i>	iii
B Alguns Endereços Interessantes	99
References	101

Capítulo 1

Introdução

A *informação* que se tem sobre uma quantidade de interesse θ é fundamental na Estatística. O verdadeiro valor de θ é desconhecido e a idéia é tentar reduzir este desconhecimento. Além disso, a intensidade da incerteza a respeito de θ pode assumir diferentes graus. Do ponto de vista Bayesiano, estes diferentes graus de incerteza são representados através de *modelos probabilísticos* para θ . Neste contexto, é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ (especificando modelos distintos). Sendo assim, não existe nenhuma distinção entre quantidades observáveis e os parâmetros de um modelo estatístico, todos são considerados quantidades aleatórias.

1.1 Teorema de Bayes

Considere uma quantidade de interesse desconhecida θ (tipicamente não observável). A informação de que dispomos sobre θ , resumida probabilisticamente através de $p(\theta)$, pode ser aumentada observando-se uma quantidade aleatória X relacionada com θ . A distribuição amostral $p(x|\theta)$ define esta relação. A idéia de que após observar $X = x$ a quantidade de informação sobre θ aumenta é bastante intuitiva e o teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação,

$$p(\theta|x) = \frac{p(x, \theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta}. \quad (1.1)$$

Note que $1/p(x)$, que não depende de θ , funciona como uma constante normalizadora de $p(\theta|x)$.

Para um valor fixo de x , a função $l(\theta; x) = p(x|\theta)$ fornece a *plausibilidade* ou *verossimilhança* de cada um dos possíveis valores de θ enquanto $p(\theta)$ é chamada distribuição *a priori* de θ . Estas duas fontes de informação, priori e verossimi-

lhança, são combinadas levando à distribuição *a posteriori* de θ , $p(\theta|x)$. Assim, a forma usual do teorema de Bayes é

$$p(\theta|x) \propto l(\theta; x)p(\theta), \quad (1.2)$$

(lê-se $p(\theta|x)$ é proporcional a $l(\theta; x)p(\theta)$). Em palavras temos que

distribuição a posteriori \propto verossimilhança \times distribuição a priori.

Note que, ao omitir o termo $p(x)$, a igualdade em (1.1) foi substituída por uma proporcionalidade. Esta forma simplificada do teorema de Bayes será útil em problemas que envolvam estimação de parâmetros já que o denominador é apenas uma constante normalizadora. Em outras situações, como seleção e comparação de modelos, este termo tem um papel crucial.

É intuitivo também que a probabilidade a posteriori de um particular conjunto de valores de θ será pequena se $p(\theta)$ ou $l(\theta; x)$ for pequena para este conjunto. Em particular, se atribuirmos probabilidade a priori igual a zero para um conjunto de valores de θ então a probabilidade a posteriori será zero qualquer que seja a amostra observada.

A partir da forma (1.2) a constante normalizadora da posteriori em (1.1) é recuperada como

$$p(x) = \int p(x, \theta)d\theta = \int p(x|\theta)p(\theta)d\theta = E_\theta[p(X|\theta)]$$

que é chamada distribuição *preditiva*. Esta é a distribuição esperada para a observação x dado θ . Assim,

- Antes de observar X podemos checar a adequação da priori fazendo previsões via $p(x)$.
- Se X observado recebia pouca probabilidade preditiva então o modelo deve ser questionado.

Em muitas aplicações (e.g. séries temporais e geoestatística) o maior interesse é na previsão do processo em pontos não observados do tempo ou espaço. Suponha então que, após observar $X = x$, estamos interessados na previsão de uma quantidade Y , também relacionada com θ , e descrita probabilisticamente por $p(y|x, \theta)$. A distribuição preditiva de Y dado x é obtida por integração como

$$p(y|x) = \int p(y, \theta|x)d\theta = \int p(y|\theta, x)p(\theta|x)d\theta. \quad (1.3)$$

Em muitos problemas estatísticos a hipótese de independência condicional entre

X e Y dado θ está presente e a distribuição preditiva fica

$$p(y|x) = \int p(y|\theta)p(\theta|x)d\theta.$$

Note no entanto que esta não é uma hipótese razoável para dados espacialmente distribuídos aonde estamos admitindo que exista alguma estrutura de correlação no espaço. De qualquer modo, em muitas aplicações práticas a integral em (1.3) não tem solução analítica e precisa ser obtida por algum método de aproximação. Note também que as previsões são sempre verificáveis uma vez que Y é uma quantidade observável. Finalmente, segue da última equação que

$$p(y|x) = E_{\theta|x}[p(Y|\theta)].$$

Fica claro também que os conceitos de *priori* e *posteriori* são relativos àquela observação que está sendo considerada no momento. Assim, $p(\theta|x)$ é a posteriori de θ em relação a X (que já foi observado) mas é a priori de θ em relação a Y (que não foi observado ainda). Após observar $Y = y$ uma nova posteriori (relativa a $X = x$ e $Y = y$) é obtida aplicando-se novamente o teorema de Bayes. Mas será que esta posteriori final depende da ordem em que as observações x e y foram processadas? Observando-se as quantidades x_1, x_2, \dots, x_n , independentes dado θ e relacionadas a θ através de $p_i(x_i|\theta)$ segue que

$$\begin{aligned} p(\theta|x_1) &\propto l_1(\theta; x_1)p(\theta) \\ p(\theta|x_2, x_1) &\propto l_2(\theta; x_2)p(\theta|x_1) \\ &\propto l_2(\theta; x_2)l_1(\theta; x_1)p(\theta) \\ &\vdots && \vdots \\ p(\theta|x_n, x_{n-1}, \dots, x_1) &\propto \left[\prod_{i=1}^n l_i(\theta; x_i) \right] p(\theta) \\ &\propto l_n(\theta; x_n)p(\theta|x_{n-1}, \dots, x_1). \end{aligned}$$

Ou seja, a ordem em que as observações são processadas pelo teorema de Bayes é irrelevante. Na verdade, elas podem até ser processadas em subgrupos.

Exemplo 1.1 : (Gamerman e Migon, 1993) Um médico, ao examinar uma pessoa, “desconfia” que ela possa ter uma certa doença. Baseado na sua experiência, no seu conhecimento sobre esta doença e nas informações dadas pelo paciente ele assume que a probabilidade do paciente ter a doença é 0,7. Aqui a quantidade

de interesse desconhecida é o indicador de doença

$$\theta = \begin{cases} 1, & \text{se o paciente tem a doença} \\ 0, & \text{se o paciente não tem a doença.} \end{cases}$$

Para aumentar sua quantidade de informação sobre a doença o médico aplica um teste X relacionado com θ através da distribuição

$$P(X = 1 \mid \theta = 0) = 0,40 \quad \text{e} \quad P(X = 1 \mid \theta = 1) = 0,95$$

e o resultado do teste foi positivo ($X = 1$).

É bem intuitivo que a probabilidade de doença deve ter aumentado após este resultado e a questão aqui é quantificar este aumento. Usando o teorema de Bayes segue que

$$P(\theta = 1 \mid X = 1) \propto l(\theta = 1; X = 1) p(\theta = 1) = (0,95)(0,7) = 0,665$$

$$P(\theta = 0 \mid X = 1) \propto l(\theta = 0; X = 1) p(\theta = 0) = (0,40)(0,3) = 0,120.$$

Uma vez que as probabilidades a posteriori somam 1, i.e.

$$P(\theta = 0 \mid X = 1) + P(\theta = 1 \mid X = 1) = 1,$$

a constante normalizadora é obtida fazendo-se $0,665k + 0,120k = 1$ e então $k = 1/0,785$. Portanto, a distribuição a posteriori de θ é

$$P(\theta = 1 \mid X = 1) = 0,665/0,785 = 0,847$$

$$P(\theta = 0 \mid X = 1) = 0,120/0,785 = 0,153.$$

O aumento na probabilidade de doença não foi muito grande porque a verossimilhança $l(\theta = 0; X = 1)$ também era grande (o modelo atribuia uma plausibilidade grande para $\theta = 0$ mesmo quando $X = 1$).

Agora o médico aplica outro teste Y cujo resultado está relacionado a θ através da seguinte distribuição

$$P(Y = 1 \mid \theta = 0) = 0,04 \quad \text{e} \quad P(Y = 1 \mid \theta = 1) = 0,99.$$

Mas antes de observar o resultado deste teste é interessante obter sua distribuição preditiva. Como θ é uma quantidade discreta segue que

$$p(y|x) = \sum_{\theta=0}^1 p(y|\theta)p(\theta|x)$$

e note que $p(\theta|x)$ é a priori em relação a Y . Assim,

$$\begin{aligned} P(Y = 1 \mid X = 1) &= P(Y = 1 \mid \theta = 0)P(\theta = 0 \mid X = 1) \\ &\quad + P(Y = 1 \mid \theta = 1)P(\theta = 1 \mid X = 1) \\ &= (0,04)(0,153) + (0,99)(0,847) = 0,845 \end{aligned}$$

$$P(Y = 0 \mid X = 1) = 1 - P(Y = 1 \mid X = 1) = 0,155.$$

O resultado deste teste foi negativo ($Y = 0$). Neste caso, é também intuitivo que a probabilidade de doença deve ter diminuído e esta redução será quantificada por uma nova aplicação do teorema de Bayes,

$$\begin{aligned} P(\theta = 1 \mid X = 1, Y = 0) &\propto l(\theta = 1; Y = 0)P(\theta = 1 \mid X = 1) \\ &\propto (0,01)(0,847) = 0,0085 \end{aligned}$$

$$\begin{aligned} P(\theta = 0 \mid X = 1, Y = 0) &\propto l(\theta = 0; Y = 0)P(\theta = 0 \mid X = 1) \\ &\propto (0,96)(0,153) = 0,1469. \end{aligned}$$

A constante normalizadora é $1/(0,0085+0,1469)=1/0,1554$ e assim a distribuição a posteriori de θ é

$$P(\theta = 1 \mid X = 1, Y = 0) = 0,0085/0,1554 = 0,055$$

$$P(\theta = 0 \mid X = 1, Y = 0) = 0,1469/0,1554 = 0,945.$$

Verifique como a probabilidade de doença se alterou ao longo do experimento

$$P(\theta = 1) = \begin{cases} 0,7, & \text{antes dos testes} \\ 0,847, & \text{após o teste } X \\ 0,055, & \text{após } X \text{ e } Y. \end{cases}$$

Note também que o valor observado de Y recebia pouca probabilidade preditiva. Isto pode levar o médico a repensar o modelo, i.e.,

- (i) Será que $P(\theta = 1) = 0,7$ é uma priori adequada?
- (ii) Será que as distribuições amostrais de X e Y estão corretas? O teste X é tão inexpressivo e Y é realmente tão poderoso?

Exemplo 1.2: Seja $Y \sim \text{Binomial}(12, \theta)$ e em um experimento observou-se $Y = 9$. A função de verossimilhança de θ é dada por

$$l(\theta) = \binom{12}{9} \theta^9 (1-\theta)^3, \quad \theta \in (0, 1).$$

Que distribuição poderia ser usada para resumir probabilisticamente nosso conhecimento sobre o parâmetro θ ? Note que, como $0 < \theta < 1$ queremos que,

$$p(\theta) = 0 \Rightarrow p(\theta|y) = 0, \quad \forall \theta \in (0, 1).$$

Podemos por exemplo assumir que $\theta \sim N(\mu, \sigma^2)$ truncada no intervalo $(0, 1)$. Neste caso, denotando por $f_N(\cdot|\mu, \sigma^2)$ a função de densidade da distribuição $N(\mu, \sigma^2)$ segue que a função de densidade a priori de θ é dada por

$$p(\theta) = \frac{f_N(\theta|\mu, \sigma^2)}{\int_0^1 f_N(\theta|\mu, \sigma^2) d\theta}.$$

Na Figura 1.1 esta função de densidade está representada para alguns valores de μ e σ^2 . Os comandos do R abaixo podem ser utilizados para gerar as curvas. Note como informações a priori bastante diferentes podem ser representadas.

```
> dnorm.t <- function(x, mean = 0, sd = 1) {
+   aux = pnorm(1, mean, sd) - pnorm(0, mean, sd)
+   dnorm(x, mean, sd)/aux
+ }
```

Outra possibilidade é através de uma reparametrização. Assumindo-se que $\delta \sim N(\mu, \sigma^2)$ e fazendo a transformação

$$\theta = \frac{\exp(\delta)}{1 + \exp(\delta)}$$

segue que a transformação inversa é simplesmente

$$\delta = \log\left(\frac{\theta}{1-\theta}\right) = \text{logito}(\theta).$$

Portanto a densidade a priori de θ fica

$$\begin{aligned} p(\theta) &= f_N(\delta(\theta)|\mu, \sigma^2) \left| \frac{d\delta}{d\theta} \right| \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\log\left(\frac{\theta}{1-\theta}\right) - \mu \right)^2 \right\} \frac{1}{\theta(1-\theta)} \end{aligned}$$

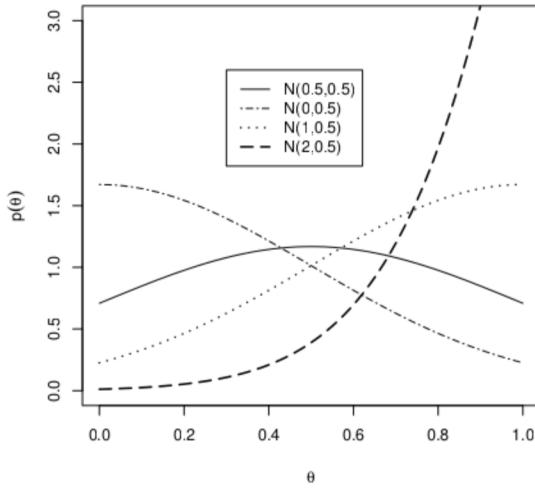


Figura 1.1: Densidades a priori normais truncadas para o parametro θ no Exemplo 1.2.

e é chamada de normal-logistica. Na Figura 1.2 esta função de densidade está representada para alguns valores de μ e σ^2 . Os comandos do R abaixo foram utilizados. Novamente note como informações a priori bastante diferentes podem ser representadas. Em particular a função de densidade de θ será sempre unimodal quando $\sigma^2 \leq 2$ e bimodal quando $\sigma^2 > 2$.

```
> dlogist = function(x, mean, sd) {
+   z = log(x/(1 - x))
+   dnorm(z, mean, sd)/(x - x^2)
+ }
```

Finalmente, podemos atribuir uma distribuição a priori $\theta \sim Beta(a, b)$ (ver Apêndice A),

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad a > 0, \quad b > 0, \quad \theta \in (0, 1).$$

Esta distribuição é simétrica em torno de 0,5 quando $a = b$ e assimétrica quando $a \neq b$. Variando os valores de a e b podemos definir uma rica família de distribuições a priori para θ , incluindo a distribuição Uniforme no intervalo (0,1) se $a = b = 1$. Algumas possibilidades estão representadas na Figura 1.3.

Um outro resultado importante ocorre quando se tem uma única observação da distribuição normal com média desconhecida. Se a média tiver priori normal

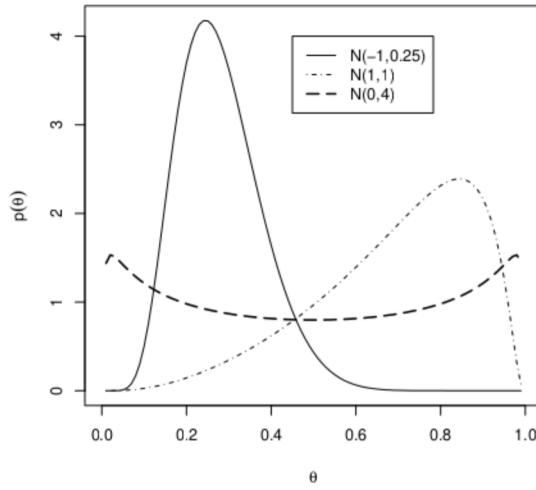


Figura 1.2: Densidades a priori tipo logísticas para o parâmetro θ no Exemplo 1.2.

então os parâmetros da posteriori são obtidos de uma forma bastante intuitiva como visto no teorema a seguir.

Teorema 1.1 Se $X|\theta \sim N(\theta, \sigma^2)$ sendo σ^2 conhecido e $\theta \sim N(\mu_0, \tau_0^{-2})$ então $\theta|x \sim N(\mu_1, \tau_1^{-2})$ sendo

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + \sigma^{-2}x}{\tau_0^{-2} + \sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau_0^{-2} + \sigma^{-2}.$$

Prova. Temos que

$$p(x|\theta) \propto \exp\{-\sigma^{-2}(x-\theta)^2/2\} \quad \text{e} \quad p(\theta) \propto \exp\{-\tau_0^{-2}(\theta-\mu_0)/2\}$$

e portanto

$$\begin{aligned} p(\theta|x) &\propto \exp\left\{-\frac{1}{2}[\sigma^{-2}(\theta^2 - 2x\theta) + \tau_0^{-2}(\theta^2 - 2\mu_0\theta)]\right\} \\ &\propto \exp\left\{-\frac{1}{2}[\theta^2(\sigma^{-2} + \tau_0^{-2}) - 2\theta(\sigma^{-2}x + \tau_0^{-2}\mu_0)]\right\}. \end{aligned}$$

sendo que os termos que não dependem de θ foram incorporados à constante de proporcionalidade. Definindo $\tau_1^{-2} = \sigma^{-2} + \tau_0^{-2}$ e $\tau_1^{-2}\mu_1 = \sigma^{-2}x - \tau_0^{-2}\mu_0$ segue que

$$p(\theta|x) \propto \exp\left\{-\frac{\tau_1^{-2}}{2}(\theta^2 - 2\theta\mu_1)\right\} \propto \exp\left\{-\frac{\tau_1^{-2}}{2}(\theta - \mu_1)^2\right\}$$

pois μ_1 não depende de θ . Portanto, a função de densidade a posteriori (a menos

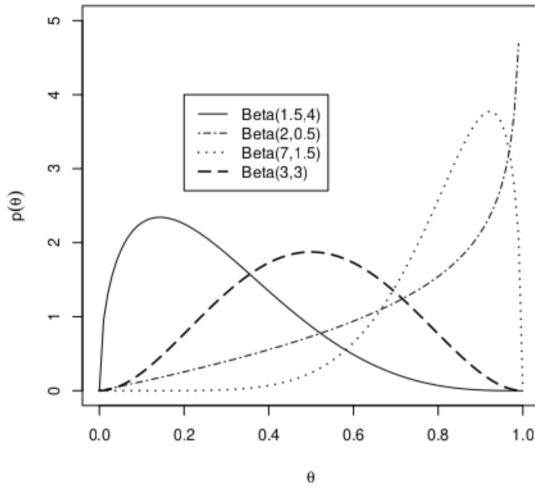


Figura 1.3: Densidades a priori Beta para o parâmetro θ no Exemplo 1.2.

de uma constante) tem a mesma forma de uma normal com média μ_1 e variância τ_1^2 .

Note que, definindo *precisão* como o inverso da variância, segue do teorema que a precisão a posteriori é a soma das precisões a priori e da verossimilhança e não depende de x . Interpretando precisão como uma medida de informação e definindo $w = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}) \in (0, 1)$ então w mede a informação relativa contida na priori com respeito à informação total. Podemos escrever então que

$$\mu_1 = w\mu_0 + (1 - w)x$$

ou seja, μ_1 é uma *combinação linear convexa* de μ_0 e x e portanto

$$\min\{\mu_0, x\} \leq \mu_1 \leq \max\{\mu_0, x\}.$$

A distribuição preditiva de X também é facilmente obtida notando que podemos reescrever as informações na forma de equações com erros não correlacionados. Assim,

$$\begin{aligned} X &= \theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \\ \theta &= \mu_0 + w, \quad w \sim N(0, \tau_0^2) \end{aligned}$$

tal que $Cov(\theta, \epsilon) = Cov(\mu_0, w) = 0$. Portanto a distribuição (incondicional) de X é normal pois ele resulta de uma soma de variáveis aleatórias com distribuição

normal. Além disso,

$$\begin{aligned} E(X) &= E(\theta) + E(\epsilon) = \mu_0 \\ Var(X) &= Var(\theta) + Var(\epsilon) = \tau_0^2 + \sigma^2 \end{aligned}$$

Conclusão, $X \sim N(\mu_0, \tau_0^2 + \sigma^2)$.

Exemplo 1.3: (Box & Tiao, 1992) Os físicos A e B desejam determinar uma constante física θ . O físico A tem mais experiência nesta área e especifica sua priori como $\theta \sim N(900, 20^2)$. O físico B tem pouca experiência e especifica uma priori muito mais incerta em relação à posição de θ , $\theta \sim N(800, 80^2)$. Assim, não é difícil verificar que

$$\text{para o físico } A: P(860 < \theta < 940) \approx 0,95$$

$$\text{para o físico } B: P(640 < \theta < 960) \approx 0,95.$$

Faz-se então uma medição X de θ em laboratório com um aparelho calibrado com distribuição amostral $X|\theta \sim N(\theta, 40^2)$ e observou-se $X = 850$. Aplicando o teorema 1.1 segue que

$$(\theta|X = 850) \sim N(890, 17, 9^2) \quad \text{para o físico } A$$

$$(\theta|X = 850) \sim N(840, 35, 7^2) \quad \text{para o físico } B.$$

Note também que os aumentos nas precisões a posteriori em relação às precisões a priori foram,

- para o físico A : precisão(θ) passou de $\tau_0^{-2} = 0,0025$ para $\tau_1^{-2} = 0,00312$ (aumento de 25%).
- para o físico B : precisão(θ) passou de $\tau_0^{-2} = 0,000156$ para $\tau_1^{-2} = 0,000781$ (aumento de 400%).

A situação está representada graficamente na Figura 1.4 a seguir. Note como a distribuição a posteriori representa um compromisso entre a distribuição a priori e a verossimilhança. Além disso, como as incertezas iniciais são bem diferentes o mesmo experimento fornece muito pouca informação adicional para o físico A enquanto que a incerteza do físico B foi bastante reduzida. Os comandos do R abaixo podem ser usados nos cálculos.

```
> norm.norm <- function(x, mu0, tau0, s0) {
+   precisao = 1/tau0 + length(x)/s0
+   tau1 = 1/precisao
+   w = (1/tau0)/precisao
+   mu1 = w * mu0 + (1 - w) * mean(x)
+   return(list(m = mu1, tau = tau1))
+ }
```

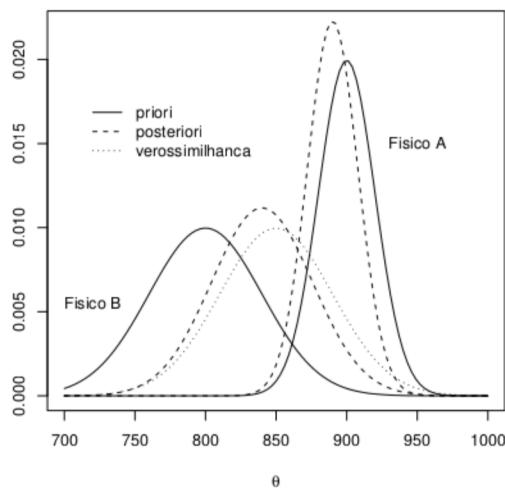


Figura 1.4: Densidades a priori e a posteriori e função de verossimilhança para o Exemplo 1.3.

1.2 Princípio da Verossimilhança

O exemplo a seguir (DeGroot, 1970, páginas 165 e 166) ilustra esta propriedade. Imagine que cada item de uma população de itens manufaturados pode ser classificado como defeituoso ou não defeituoso. A proporção θ de itens defeituosos na população é desconhecida e uma amostra de itens será selecionada de acordo com um dos seguintes métodos:

- (i) n itens serão selecionados ao acaso.
- (ii) Itens serão selecionados ao acaso até que y defeituosos sejam obtidos.
- (iii) Itens serão selecionados ao acaso até que o inspetor seja chamado para resolver um outro problema.

- (iv) Itens serão selecionados ao acaso até que o inspetor decida que já acumulou informação suficiente sobre θ .

Qualquer que tenha sido o esquema amostral, se foram inspecionados n itens x_1, \dots, x_n dos quais y eram defeituosos então

$$l(\theta; x) \propto \theta^y (1 - \theta)^{n-y}.$$

O Princípio da Verossimilhança postula que para fazer inferência sobre uma quantidade de interesse θ só importa aquilo que foi realmente observado e não aquilo que “poderia” ter ocorrido mas efetivamente não ocorreu.

1.3 Exercícios

1. No Exemplo 1.3, obtenha também a distribuição preditiva de X e compare o valor observado com a média desta preditiva para os 2 físicos. Faça uma previsão para uma 2^a medição Y feita com o mesmo aparelho.
2. Uma máquina produz 5% de itens defeituosos. Cada item produzido passa por um teste de qualidade que o classifica como “bom”, “defeituoso” ou “suspeito”. Este teste classifica 20% dos itens defeituosos como bons e 30% como suspeitos. Ele também classifica 15% dos itens bons como defeituosos e 25% como suspeitos.
 - (a) Que proporção dos itens serão classificados como suspeitos ?
 - (b) Qual a probabilidade de um item classificado como suspeito ser defeituoso ?
 - (c) Outro teste, que classifica 95% dos itens defeituosos e 1% dos itens bons como defeituosos, é aplicado somente aos itens suspeitos.
 - (d) Que proporção de itens terão a suspeita de defeito confirmada ?
 - (e) Qual a probabilidade de um item reprovado neste 2^a teste ser defeituoso ?
3. Uma empresa de crédito precisa saber como a inadimplência está distribuída entre seus clientes. Sabe-se que um cliente pode pertencer às classes A , B , C ou D com probabilidades 0,50, 0,20, 0,20 e 0,10 respectivamente. Um cliente da classe A tem probabilidade 0,30 de estar inadimplente, um da classe B tem probabilidade 0,10 de estar inadimplente, um da classe C tem probabilidade 0,05 de estar inadimplente e um da classe D tem probabilidade 0,05 de estar inadimplente. Um cliente é sorteado aleatoriamente.
 - (a) Defina os eventos e enumere as probabilidades fornecidas no problema.

- (b) Qual a probabilidade dele estar inadimplente ?
- (c) Sabendo que ele está inadimplente, qual a sua classe mais provável?
4. Suponha que seus dados x_1, \dots, x_n são processados sequencialmente, i.e. x_1 é observado antes de x_2 e assim por diante. Escreva um programa que aplica o Teorema 1.1 obtendo a média e a variância a posteriori dado x_1 , use esta distribuição como priori para obter a média e a variância a posteriori dados x_1, x_2 e repita o procedimento sequencialmente até obter a posteriori dados x_1, \dots, x_n . Faça um gráfico com as médias a posteriori mais ou menos 2 desvios padrão a posteriori.

Capítulo 2

Distribuições a Priori

A utilização de informação a priori em inferência Bayesiana requer a especificação de uma distribuição a priori para a quantidade de interesse θ . Esta distribuição deve representar (probabilisticamente) o conhecimento que se tem sobre θ antes da realização do experimento. Neste capítulo serão discutidas diferentes formas de especificação da distribuição a priori.

2.1 Prioris Conjugadas

A partir do conhecimento que se tem sobre θ , pode-se definir uma família paramétrica de densidades. Neste caso, a distribuição a priori é representada por uma forma funcional, cujos parâmetros devem ser especificados de acordo com este conhecimento. Estes parâmetros indexadores da família de distribuições a priori são chamados de *hiperparâmetros* para distingui-los dos parâmetros de interesse θ .

Esta abordagem em geral facilita a análise e o caso mais importante é o de prioris conjugadas. A idéia é que as distribuições a priori e a posteriori pertençam à mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método Bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas.

Definição 2.1 Se $F = \{p(x|\theta), \theta \in \Theta\}$ é uma classe de distribuições amostrais então uma classe de distribuições P é conjugada a F se

$$\forall p(x|\theta) \in F \quad e \quad p(\theta) \in P \Rightarrow p(\theta|x) \in P.$$

Gamerman (1996, 1997 Cap. 2) alerta para o cuidado com a utilização indiscriminada de prioris conjugadas. Essencialmente, o problema é que a priori

conjugada nem sempre é uma representação adequada da incerteza a priori. Sua utilização está muitas vezes associada à tratabilidade analítica decorrente.

Uma vez entendidas suas vantagens e desvantagens a questão que se coloca agora é “como” obter uma família de distribuições conjugadas.

- (i) Identifique a classe P de distribuições para θ tal que $l(\theta; x)$ seja proporcional a um membro desta classe.
- (ii) Verifique se P é *fechada por amostragem*, i.e., se $\forall p_1, p_2 \in P \exists k$ tal que $kp_1p_2 \in P$.

Se, além disso, existe uma constante k tal que $k^{-1} = \int l(\theta; x)d\theta < \infty$ e todo $p \in P$ é definido como $p(\theta) = k l(\theta; x)$ então P é a *família conjugada natural* ao modelo amostral gerador de $l(\theta; x)$.

Exemplo 2.1: Sejam $X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$. Então a densidade amostral conjunta é

$$p(\mathbf{x}|\theta) = \theta^t(1-\theta)^{n-t}, \quad 0 < \theta < 1 \quad \text{sendo} \quad t = \sum_{i=1}^n x_i$$

e pelo teorema de Bayes segue que

$$p(\theta|\mathbf{x}) \propto \theta^t(1-\theta)^{n-t}p(\theta).$$

Note que $l(\theta; x)$ é proporcional à densidade de uma distribuição Beta($t + 1, n - t + 1$). Além disso, se p_1 e p_2 são as densidades das distribuições Beta(a_1, b_1) e Beta(a_2, b_2) então

$$p_1p_2 \propto \theta^{a_1+a_2-2}(1-\theta)^{b_1+b_2-2},$$

ou seja p_1p_2 é proporcional a densidade da distribuição Beta($a_1 + a_2 - 1, b_1 + b_2 - 1$). Conclui-se que a família de distribuições Beta com parâmetros inteiros é conjugada natural à família Bernoulli. Na prática esta classe pode ser ampliada para incluir todas as distribuições Beta, i.e. incluindo todos os valores positivos dos parâmetros.

2.2 Conjugaçāo na Família Exponencial

A família exponencial inclui muitas das distribuições de probabilidade mais comumente utilizadas em Estatística, tanto contínuas quanto discretas. Uma característica essencial desta família é que existe uma estatística suficiente com dimensão

fixa. Veremos adiante que a classe conjugada de distribuições é muito fácil de caracterizar.

Definição 2.2 *A família de distribuições com função de (densidade) de probabilidade $p(x|\theta)$ pertence à família exponencial a um parâmetro se podemos escrever*

$$p(x|\theta) = a(x) \exp\{u(x)\phi(\theta) + b(\theta)\}.$$

Note que pelo critério de fatoração de Neyman $U(x)$ é uma estatística suficiente para θ .

Neste caso, a classe conjugada é facilmente identificada como,

$$p(\theta) = k(\alpha, \beta) \exp\{\alpha\phi(\theta) + \beta b(\theta)\}.$$

e aplicando o teorema de Bayes segue que

$$p(\theta|x) = k(\alpha + u(x), \beta + 1) \exp\{[\alpha + u(x)]\phi(\theta) + [\beta + 1]b(\theta)\}.$$

Agora, usando a constante k , a distribuição preditiva pode ser facilmente obtida sem necessidade de qualquer integração. A partir da equação $p(x)p(\theta|x) = p(x|\theta)p(\theta)$ e após alguma simplificação segue que

$$p(x) = \frac{p(x|\theta)p(\theta)}{p(\theta|x)} = \frac{a(x)k(\alpha, \beta)}{k(\alpha + u(x), \beta + 1)}.$$

Exemplo 2.2: Uma extensão direta do Exemplo 2.1 é o modelo binomial, i.e. $X|\theta \sim \text{Binomial}(n, \theta)$. Neste caso,

$$p(x|\theta) = \binom{n}{x} \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + n \log(1-\theta)\right\}$$

e a família conjugada natural é Beta(r, s). Podemos escrever então

$$\begin{aligned} p(\theta) &\propto \theta^{r-1}(1-\theta)^{s-1} \\ &\propto \exp\left\{(r-1)\log\left(\frac{\theta}{1-\theta}\right) + \left(\frac{s+r-2}{n}\right)n \log(1-\theta)\right\} \\ &\propto \exp\{\alpha\phi(\theta) + \beta b(\theta)\}. \end{aligned}$$

A posteriori também é Beta com parâmetros $\alpha + x$ e $\beta + 1$ ou equivalentemente

$r + x$ e $s + n - x$, i.e.

$$\begin{aligned} p(\theta|x) &\propto \exp \left\{ (r+x-1)\phi(\theta) + \left[\frac{s+r-2+n}{n} \right] b(\theta) \right\} \\ &\propto \theta^{r+x-1} (1-\theta)^{s+n-x-1}. \end{aligned}$$

Como ilustração, no Exemplo 2.2 suponha que $n = 12$, $X = 9$ e usamos prioris conjugadas Beta(1,1), Beta(2,2) e Beta(1,3). As funções de densidade destas distribuições juntamente com a função de verossimilhança normalizada e as respectivas densidades a posteriori estão na Figura 2.1. A distribuição preditiva é dada por

$$p(x) = \binom{n}{x} \frac{B(r+x, s+n-x)}{B(r, s)}, \quad x = 0, 1, \dots, n, \quad n \geq 1,$$

onde B^{-1} é a constante normalizadora da distribuição Beta, i.e. (ver Apêndice A)

$$B^{-1}(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}.$$

Esta distribuição é denominada Beta-Binomial.

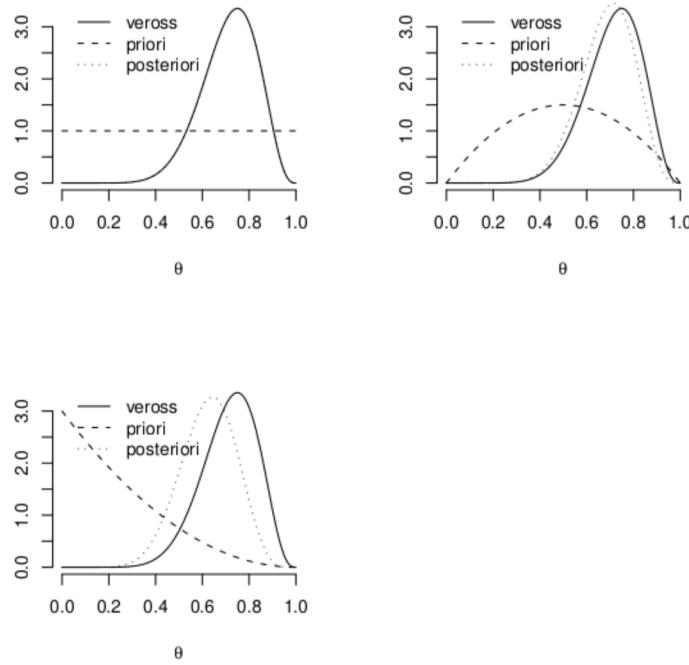


Figura 2.1: Densidades a priori, a posteriori e função de verossimilhança normalizada para o Exemplo 2.2.

No Exemplo 2.2 suponha novamente que $n = 12$, $X = 9$ e usamos as prioris conjugadas Beta(1,1), Beta(2,2) e Beta(1,3). Na Tabela 2.1 estão listadas as probabilidades preditivas $P(X = k)$ associadas a estas prioris. Os comandos do R a seguir podem ser usados no cálculo destas probabilidades.

```
> beta.binomial = function(n, a, b) {
+   m = matrix(0, n + 1, 2)
+   m[, 1] = 0:n
+   for (x in 0:n) m[x, 2] = round(choose(n, x) * beta(a + x,
+     b + n - x)/beta(a, b), 4)
+   return(list(m = m))
+ }
```

Tabela 2.1: Probabilidades preditivas da Beta-Binomial para o Exemplo 2.2

k	Beta(1,1)	Beta(2,2)	Beta(1,3)
0	0.0769	0.0527	0.1714
1	0.0769	0.0725	0.1451
2	0.0769	0.0879	0.1209
3	0.0769	0.0989	0.0989
4	0.0769	0.1055	0.0791
5	0.0769	0.1077	0.0615
6	0.0769	0.1055	0.0462
7	0.0769	0.0989	0.0330
8	0.0769	0.0879	0.0220
9	0.0769	0.0725	0.0132
10	0.0769	0.0527	0.0066
11	0.0769	0.0286	0.0022
12	0.0000	0.0000	0.0000

No caso geral em que se tem uma amostra X_1, \dots, X_n da família exponencial a natureza sequencial do teorema de Bayes permite que a análise seja feita por replicações sucessivas. Assim a cada observação x_i os parâmetros da distribuição a posteriori são atualizados via

$$\begin{aligned}\alpha_i &= \alpha_{i-1} + u(x_i) \\ \beta_i &= \beta_{i-1} + 1\end{aligned}$$

com $\alpha_0 = \alpha$ e $\beta_0 = \beta$. Após n observações temos que

$$\begin{aligned}\alpha_n &= \alpha + \sum_{i=1}^n u(x_i) \\ \beta_n &= \beta + n\end{aligned}$$

e a distribuição preditiva é dada por

$$p(\mathbf{x}) = \left[\prod_{i=1}^n a(x_i) \right] \frac{k(\alpha, \beta)}{k(\alpha + \sum u(x_i), \beta + n)}.$$

Finalmente, a definição de família exponencial pode ser extendida ao caso multiparamétrico, i.e.

$$p(\mathbf{x}|\boldsymbol{\theta}) = \left[\prod_{i=1}^n a(x_i) \right] \exp \left\{ \sum_{j=1}^r \left[\sum_{i=1}^n u_j(x_i) \right] \phi_j(\boldsymbol{\theta}) + nb(\boldsymbol{\theta}) \right\}$$

com $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$. Neste caso, pelo critério de fatoração, temos que $\sum U_1(x_i), \dots, \sum U_r(x_i)$ é uma estatística conjuntamente suficiente para o vetor de parâmetros $\boldsymbol{\theta}$.

2.3 Principais Famílias Conjugadas

Já vimos que a família de distribuições Beta é conjugada ao modelo Bernoulli e binomial. Não é difícil mostrar que o mesmo vale para as distribuições amostrais geométrica e binomial-negativa (ver Exercício 1). A seguir veremos resultados para outros membros importantes da família exponencial.

2.3.1 Distribuição normal com variância conhecida

Para uma única observação vimos pelo Teorema 1.1 que a família de distribuições normais é conjugada ao modelo normal. Para uma amostra de tamanho n , a função de verossimilhança pode ser escrita como

$$\begin{aligned}l(\theta; \mathbf{x}) &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right\} \\ &\propto \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \theta)^2 \right\}\end{aligned}$$

onde os termos que não dependem de θ foram incorporados à constante de proporcionalidade. Portanto, a verossimilhança tem a mesma forma daquela baseada em uma única observação bastando substituir x por \bar{x} e σ^2 por σ^2/n . Logo vale

o Teorema 1.1 com as devidas substituições, i.e. a distribuição a posteriori de θ dado \mathbf{x} é $N(\mu_1, \tau_1^2)$ sendo

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\bar{x}}{\tau_0^{-2} + n\sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}.$$

Note que a média a posteriori pode ser reescrita como $w\mu_0 + (1 - w)\bar{x}$ sendo $w = \tau_0^{-2}/(\tau_0^{-2} + n\sigma^{-2})$.

Uma função geral pode ser escrita no R para calcular estes parâmetros e opcionalmente fazer os gráficos das densidades.

```
> norm.norm <- function(x, sigma, mu0, tau0, plot = F) {
+   n = length(x)
+   xbar = mean(x)
+   ep = sigma/sqrt(n)
+   sigma2 = sigma^2
+   tau1 = n * (1/sigma2) + (1/tau0)
+   mu1 = (n * (1/sigma2) * xbar + (1/tau0) * mu0)/tau1
+   if (plot) {
+     curve(dnorm(x, xbar, ep), xbar - 3 * ep, xbar + 3 * ep)
+     curve(dnorm(x, mu0, sqrt(tau0)), add = T, col = 2)
+     curve(dnorm(x, mu1, 1/sqrt(tau1)), add = T, col = 3)
+     legend(-0.5, 1.2, legend = c("veross.", "priori", "posteriori"),
+            col = 1:3, lty = c(1, 1, 1))
+   }
+   return(list(mu1 = mu1, tau1 = tau1))
+ }
```

2.3.2 Distribuição de Poisson

Seja X_1, \dots, X_n uma amostra aleatória da distribuição de Poisson com parâmetro θ . Sua função de probabilidade conjunta é dada por

$$p(x|\theta) = \frac{e^{-n\theta}\theta^t}{\prod x_i!} \propto e^{-n\theta}\theta^t, \quad \theta > 0, \quad t = \sum_{i=1}^n x_i.$$

O núcleo da verossimilhança é da forma $\theta^a e^{-b\theta}$ que caracteriza a família de distribuições Gama a qual é fechada por amostragem (verifique!). Assim, distribuição a priori conjugada natural de θ é Gama com parâmetros positivos α e β , i.e.

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \quad \alpha > 0, \quad \beta > 0, \quad \theta > 0.$$

A densidade a posteriori fica

$$p(\theta|x) \propto \theta^{\alpha+t-1} \exp\{-(\beta+n)\theta\}$$

que corresponde à densidade Gama($\alpha + t, \beta + n$). Note que a média a posteriori pode ser reescrita como uma combinação linear da média a priori e da média amostral (ver exercício 6). A distribuição preditiva também é facilmente obtida pois

$$p(x|\theta) = \left[\prod_{i=1}^n \frac{1}{x_i!} \right] \exp\{t \log \theta - n\theta\}$$

e portanto

$$p(x) = \left[\prod_{i=1}^n \frac{1}{x_i!} \right] \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+t)}{(\beta+n)^{\alpha+t}}.$$

Para uma única observação x segue então que

$$\begin{aligned} p(x) &= \frac{1}{x!} \frac{\beta^\alpha \Gamma(\alpha+x)}{\Gamma(\alpha) (\beta+1)^{\alpha+x}} = \frac{1}{x!} \left(\frac{\beta}{\beta+1} \right)^\alpha \left(\frac{1}{\beta+1} \right)^x \frac{(\alpha+x-1)!}{(\alpha-1)!} \\ &= \binom{\alpha+x-1}{x} \left(\frac{\beta}{\beta+1} \right)^\alpha \left(\frac{1}{\beta+1} \right)^x. \end{aligned}$$

Esta distribuição é chamada de Binomial-Negativa com parâmetros α e β e sua média e variância são facilmente obtidos como

$$E(X) = E[E(X|\theta)] = E(\theta) = \alpha/\beta$$

$$Var(X) = E[Var(X|\theta)] + Var[E(X|\theta)] = E(\theta) + Var(\theta) = \frac{\alpha(\beta+1)}{\beta^2}.$$

2.3.3 Distribuição multinomial

Denotando por $\mathbf{X} = (X_1, \dots, X_p)$ o número de ocorrências em cada uma de p categorias em n ensaios independentes e por $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ as probabilidades associadas, deseja-se fazer inferência sobre estes p parâmetros. No entanto, note que existem efetivamente $p-1$ parâmetros já que temos a seguinte restrição $\sum_{i=1}^p \theta_i = 1$. Além disso, a restrição $\sum_{i=1}^p X_i = n$ obviamente também se aplica. Dizemos que \mathbf{X} tem distribuição multinomial com parâmetros n e $\boldsymbol{\theta}$ e função de probabilidade conjunta das p contagens \mathbf{X} é dada por

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^p x_i!} \prod_{i=1}^p \theta_i^{x_i}.$$

Note que esta é uma generalização da distribuição binomial que tem apenas duas categorias. Não é difícil mostrar que esta distribuição também pertence à família exponencial. A função de verossimilhança para $\boldsymbol{\theta}$ é

$$l(\boldsymbol{\theta}; \mathbf{x}) \propto \prod_{i=1}^p \theta_i^{x_i}$$

que tem o mesmo núcleo da função de densidade de uma distribuição de Dirichlet. A família Dirichlet com parâmetros inteiros a_1, \dots, a_p é a conjugada natural do modelo multinomial, porém na prática a conjugação é extendida para parâmetros não inteiros. A distribuição a posteriori é dada por

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto \prod_{i=1}^p \theta_i^{x_i} \prod_{i=1}^p \theta_i^{a_i-1} = \prod_{i=1}^p \theta_i^{x_i+a_i-1}.$$

Note que estamos generalizando a análise conjugada para amostras binomiais com priori beta.

2.3.4 Distribuição normal com média conhecida e variância desconhecida

Seja X_1, \dots, X_n uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ conhecido e $\phi = \sigma^{-2}$ desconhecido. Neste caso a função de densidade conjunta é dada por

$$p(\mathbf{x}|\theta, \phi) \propto \phi^{n/2} \exp\left\{-\frac{\phi}{2} \sum_{i=1}^n (x_i - \theta)^2\right\}.$$

Note que o núcleo desta verossimilhança tem a mesma forma daquele de uma distribuição Gama. Como sabemos que a família Gama é fechada por amostragem podemos considerar uma distribuição a priori Gama com parâmetros $n_0/2$ e $n_0\sigma_0^2/2$, i.e.

$$\phi \sim Gama\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Equivalentemente, podemos atribuir uma distribuição a priori qui-quadrado com n_0 graus de liberdade para $n_0\sigma_0^2\phi$. A forma funcional dos parâmetros da distribuição a priori é apenas uma conveniência matemática como veremos a seguir.

Definindo $ns_0^2 = \sum_{i=1}^n (x_i - \theta)^2$ e aplicando o teorema de Bayes obtemos a

distribuição a posteriori de ϕ ,

$$\begin{aligned} p(\phi|\mathbf{x}) &\propto \phi^{n/2} \exp\left\{-\frac{\phi}{2}ns_0^2\right\} \phi^{n_0/2-1} \exp\left\{-\frac{\phi}{2}n_0\sigma_0^2\right\} \\ &= \phi^{(n_0+n)/2-1} \exp\left\{-\frac{\phi}{2}(n_0\sigma_0^2 + ns_0^2)\right\}. \end{aligned}$$

Note que esta expressão corresponde ao núcleo da distribuição Gama, como era esperado devido à conjugação. Portanto,

$$\phi|\mathbf{x} \sim \text{Gama}\left(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2 + ns_0^2}{2}\right).$$

Equivalentemente podemos dizer que $(n_0\sigma_0^2 + ns_0^2)\phi \mid \mathbf{x} \sim \chi_{n_0+n}^2$.

2.3.5 Distribuição normal com média e variância desconhecidos

Seja X_1, \dots, X_n uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com ambos θ e $\phi = \sigma^{-2}$ desconhecidos. Precisamos então especificar uma distribuição a priori conjunta para θ e ϕ . Uma possibilidade é fazer a especificação em dois estágios já que podemos sempre escrever $p(\theta, \phi) = p(\theta|\phi)p(\phi)$. No primeiro estágio,

$$\theta|\phi \sim N(\mu_0, (c_0\phi)^{-1}), \quad \phi = \sigma^{-2}$$

e a distribuição a priori marginal de ϕ é a mesma do caso anterior, i.e.

$$\phi \sim \text{Gama}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

A distribuição conjunta de (θ, ϕ) é geralmente chamada de Normal-Gama com parâmetros $(\mu_0, c_0, n_0, \sigma_0^2)$ e sua função de densidade conjunta é dada por,

$$\begin{aligned} p(\theta, \phi) &= p(\theta|\phi)p(\phi) \\ &\propto \phi^{1/2} \exp\left\{-\frac{c_0\phi}{2}(\theta - \mu_0)^2\right\} \phi^{n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2\phi}{2}\right\} \\ &\propto \phi^{(n_0+1)/2-1} \exp\left\{-\frac{\phi}{2}(n_0\sigma_0^2 + c_0(\theta - \mu_0)^2)\right\}. \end{aligned}$$

A partir desta densidade conjunta podemos obter a distribuição marginal de

θ por integração

$$\begin{aligned} p(\theta) &= \int p(\theta|\phi)p(\phi)d\phi \\ &\propto \int_0^\infty \phi^{(n_0+1)/2-1} \exp\left\{-\frac{\phi}{2}[n_0\sigma_0^2 + c_0(\theta - \mu_0)^2]\right\} d\phi \\ &\propto \left[\frac{n_0\sigma_0^2 + c_0(\theta - \mu_0)^2}{2}\right]^{-\frac{n_0+1}{2}} \propto \left[1 + \frac{(\theta - \mu_0)^2}{n_0(\sigma_0^2/c_0)}\right]^{-\frac{n_0+1}{2}}, \end{aligned}$$

que é o núcleo da distribuição t de Student com n_0 graus de liberdade, parâmetro de locação μ_0 e parâmetro de escala σ_0^2/c_0 (ver Apêndice A). Denotamos $\theta \sim t_{n_0}(\mu_0, \sigma_0^2/c_0)$. A distribuição condicional de ϕ dado θ também é facilmente obtida como

$$\begin{aligned} p(\phi|\theta) &\propto p(\theta|\phi)p(\phi) \\ &\propto \phi^{(n_0+1)/2-1} \exp\left\{-\frac{\phi}{2}[n_0\sigma_0^2 + c_0(\theta - \mu_0)^2]\right\}, \end{aligned}$$

e portanto,

$$\phi|\theta \sim \text{Gama}\left(\frac{n_0 + 1}{2}, \frac{n_0\sigma_0^2 + c_0(\theta - \mu_0)^2}{2}\right).$$

A posteriori conjunta de (θ, ϕ) também é obtida em 2 etapas como segue. Primeiro, para ϕ fixo podemos usar o resultado da Seção 2.3.1 de modo que a distribuição a posteriori de θ dado ϕ fica

$$\theta|\phi, \mathbf{x} \sim N(\mu_1, (c_1\phi)^{-1})$$

sendo

$$\mu_1 = \frac{c_0\phi\mu_0 + n\phi\bar{x}}{c_0\phi + n\phi} = \frac{c_0\mu_0 + n\bar{x}}{c_0 + n} \quad \text{e} \quad c_1 = c_0 + n.$$

Na segunda etapa, combinando a verossimilhança com a priori de ϕ obtemos que

$$\phi|\mathbf{x} \sim \text{Gama}\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right)$$

sendo

$$n_1 = n_0 + n \quad \text{e} \quad n_1\sigma_1^2 = n_0\sigma_0^2 + \sum(x_i - \bar{x})^2 + c_0n(\mu_0 - \bar{x})^2/(c_0 + n).$$

Equivalentemente, podemos escrever a posteriori de ϕ como $n_1\sigma_1^2\phi \sim \chi_{n_1}^2$. Assim, a posteriori conjunta é $(\theta, \phi|\mathbf{x}) \sim \text{Normal-Gama}(\mu_1, c_1, n_1, \sigma_1^2)$ e portanto a

posteriori marginal de θ fica

$$\theta \mid \mathbf{x} \sim t_{n_1}(\mu_1, \sigma_1^2/c_1).$$

Em muitas situações é mais fácil pensar em termos de algumas características da distribuição a priori do que em termos de seus hiperparâmetros. Por exemplo, se $E(\theta) = 2$, $Var(\theta) = 5$, $E(\phi) = 3$ e $Var(\phi) = 3$ então

(i) $\mu_0 = 2$ pois $E(\theta) = \mu_0$.

(ii) $\sigma_0^2 = 1/3$ pois $E(\phi) = 1/\sigma_0^2$.

(iii) $n_0 = 6$ pois $Var(\phi) = 2/(n_0\sigma_0^4) = 18/n_0$.

(iv) $c_0 = 1/10$ pois $Var(\theta) = \left(\frac{n_0}{n_0 - 2}\right) \frac{\sigma_0^2}{c_0} = \frac{1}{2c_0}$

2.4 Priori não Informativa

Esta seção refere-se a especificação de distribuições a priori quando se espera que a informação dos dados seja dominante, no sentido de que a nossa informação a priori é *vaga*. Os conceitos de “conhecimento vago”, “não informação”, ou “ignorância a priori” claramente não são únicos e o problema de caracterizar prioris com tais características pode se tornar bastante complexo.

Por outro lado, reconhece-se a necessidade de alguma forma de análise que, em algum sentido, consiga captar esta noção de uma priori que tenha um efeito mínimo, relativamente aos dados, na inferência final. Tal análise pode ser pensada como um ponto de partida quando não se consegue fazer uma elicitação detalhada do “verdadeiro” conhecimento a priori. Neste sentido, serão apresentadas aqui algumas formas de “como” fazer enquanto discussões mais detalhadas são encontradas em Berger (1985), Box & Tiao (1992), Bernardo & Smith (1994) e O’Hagan (1994).

A primeira idéia de “não informação” a priori que se pode ter é pensar em todos os possíveis valores de θ como igualmente prováveis, i.e. com uma distribuição a priori uniforme. Neste caso, fazendo $p(\theta) \propto k$ para θ variando em um subconjunto da reta significa que nenhum valor particular tem preferência (Bayes, 1763). Porém esta escolha de priori pode trazer algumas dificuldades técnicas,

- (i) Se o intervalo de variação de θ for ilimitado então a distribuição a priori é imprópria, i.e.

$$\int p(\theta) d\theta = \infty.$$

- (ii) Se $\phi = g(\theta)$ é uma reparametrização não linear monótona de θ então $p(\phi)$ é não uniforme já que pelo teorema de transformação de variáveis

$$p(\phi) = p(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right|.$$

Na prática, como estaremos interessados na distribuição a posteriori não daremos muita importância à impropriedade da distribuição a priori. No entanto devemos sempre nos certificar de que a posterior é própria antes de fazer qualquer inferência.

A classe de prioris não informativas proposta por Jeffreys (1961) é invariante a transformações 1 a 1, embora em geral seja imprópria e será definida a seguir. Antes porém precisamos da definição da medida de informação de Fisher.

Definição 2.3 Considerando uma única observação X com função de (densidade) de probabilidade $p(x|\theta)$. A medida de informação esperada de Fisher de θ através de X é definida como

$$I(\theta) = E \left[-\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \right].$$

Se $\boldsymbol{\theta}$ for um vetor paramétrico define-se então a matriz de informação esperada de Fisher de $\boldsymbol{\theta}$ através de X como

$$\mathbf{I}(\boldsymbol{\theta}) = E \left[-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right].$$

Note que o conceito de informação aqui está sendo associado a uma espécie de curvatura média da função de verossimilhança no sentido de que quanto maior a curvatura mais precisa é a informação contida na verossimilhança, ou equivalente maior o valor de $I(\theta)$. Em geral espera-se que a curvatura seja negativa e por isso seu valor é tomado com sinal trocado. Note também que a esperança matemática é tomada em relação à distribuição amostral $p(x|\theta)$.

Podemos considerar então $I(\theta)$ uma medida de informação global enquanto que uma medida de informação local é obtida quando não se toma o valor esperado na definição acima. A medida de informação observada de Fisher $J(\theta)$ fica então definida como

$$J(\theta) = -\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}$$

e que será utilizada mais adiante quando falarmos sobre estimação.

Definição 2.4 Seja uma observação X com função de (densidade) de probabilidade $p(x|\theta)$. A priori não informativa de Jeffreys tem função de densidade dada por

$$p(\theta) \propto [I(\theta)]^{1/2}.$$

Se $\boldsymbol{\theta}$ for um vetor paramétrico então $p(\boldsymbol{\theta}) \propto |\det \mathbf{I}(\boldsymbol{\theta})|^{1/2}$.

Exemplo 2.3: Seja $X_1, \dots, X_n \sim \text{Poisson}(\theta)$. Então o logaritmo da função de probabilidade conjunta é dado por

$$\log p(\mathbf{x}|\theta) = -n\theta + \sum_{i=1}^n x_i \log \theta - \log \prod_{i=1}^n x_i!$$

e tomado-se a segunda derivada segue que

$$\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta} \left[-n + \frac{\sum_{i=1}^n x_i}{\theta} \right] = -\frac{\sum_{i=1}^n x_i}{\theta^2}$$

e assim,

$$I(\theta) = \frac{1}{\theta^2} E \left[\sum_{i=1}^n x_i \right] = n/\theta \propto \theta^{-1}.$$

Portanto, a priori não informativa de Jeffreys para θ no modelo Poisson é $p(\theta) \propto \theta^{-1/2}$. Note que esta priori é obtida tomado-se a conjugada natural Gama(α, β) e fazendo-se $\alpha = 1/2$ e $\beta \rightarrow 0$.

Em geral a priori não informativa é obtida fazendo-se o parâmetro de escala da distribuição conjugada tender a zero e fixando-se os demais parâmetros convenientemente. Além disso, a priori de Jeffreys assume formas específicas em alguns modelos que são frequentemente utilizados como veremos a seguir.

Definição 2.5 *X tem um modelo de locação se existem uma função f e uma quantidade θ tais que $p(x|\theta) = f(x - \theta)$. Neste caso θ é chamado de parâmetro de locação.*

A definição vale também quando θ é um vetor de parâmetros. Alguns exemplos importantes são a distribuição normal com variância conhecida, e a distribuição normal multivariada com matriz de variância-covariância conhecida. Pode-se mostrar que para o modelo de locação a priori de Jeffreys é dada por $p(\theta) \propto \text{constante}$.

Definição 2.6 *X tem um modelo de escala se existem uma função f e uma quantidade σ tais que $p(x|\sigma) = (1/\sigma)f(x/\sigma)$. Neste caso σ é chamado de parâmetro de escala.*

Alguns exemplos são a distribuição exponencial com parâmetro θ , com parâmetro de escala $\sigma = 1/\theta$, e a distribuição $N(\theta, \sigma^2)$ com média conhecida e escala σ . Pode-se mostrar que para o modelo de escala a priori de Jeffreys é dada por $p(\sigma) \propto \sigma^{-1}$.

Definição 2.7 X tem um modelo de locação e escala se existem uma função f e as quantidades θ e σ tais que

$$p(x|\theta, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\theta}{\sigma}\right).$$

Neste caso θ é chamado de parâmetro de locação e σ de parâmetro de escala.

Alguns exemplos são a distribuição normal (uni e multivariada) e a distribuição de Cauchy. Em modelos de locação e escala, a priori não informativa pode ser obtida assumindo-se independência a priori entre θ e σ de modo que $p(\theta, \sigma) = p(\theta)p(\sigma) \propto \sigma^{-1}$.

Exemplo 2.4: Seja $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ com μ e σ^2 desconhecidos. Neste caso,

$$p(x|\mu, \sigma^2) \propto \frac{1}{\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

portanto (μ, σ) é parâmetro de locação-escala e $p(\mu, \sigma) \propto \sigma^{-1}$ é a priori não informativa. Então, pela propriedade da invariância, a priori não informativa para (μ, σ^2) no modelo normal é $p(\mu, \sigma^2) \propto \sigma^{-2}$.

Vale notar entretanto que a priori não informativa de Jeffreys viola o princípio da verossimilhança, já que a informação de Fisher depende da distribuição amostral.

2.5 Prioris Hierárquicas

A idéia aqui é dividir a especificação da distribuição a priori em estágios. Além de facilitar a especificação esta abordagem é natural em determinadas situações experimentais.

A distribuição a priori de θ depende dos valores dos hiperparâmetros ϕ e podemos escrever $p(\theta|\phi)$ ao invés de $p(\theta)$. Além disso, ao invés de fixar valores para os hiperparâmetros podemos especificar uma distribuição a priori $p(\phi)$ completando assim o segundo estágio na hierarquia. Assim, a distribuição a priori conjunta é simplesmente $p(\theta, \phi) = p(\theta|\phi)p(\phi)$ e a distribuição a priori marginal de θ pode ser então obtida por integração como

$$p(\theta) = \int p(\theta, \phi)d\phi = \int p(\theta|\phi)p(\phi)d\phi.$$

A distribuição a posteriori conjunta fica

$$p(\theta, \phi | \mathbf{x}) \propto p(\mathbf{x} | \theta, \phi) p(\theta | \phi) p(\phi)$$

pois a distribuição dos dados depende somente de θ . Em outras palavras, dado θ , \mathbf{x} e ϕ são independentes.

Exemplo 2.5 : Sejam X_1, \dots, X_n tais que $X_i \sim N(\theta_i, \sigma^2)$ com σ^2 conhecido e queremos especificar uma distribuição a priori para o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$. Suponha que no primeiro estágio assumimos que $\theta_i \sim N(\mu, \tau^2)$, $i = 1, \dots, n$. Neste caso, se fixarmos o valor de $\tau^2 = \tau_0^2$ e assumirmos que μ tem distribuição normal então $\boldsymbol{\theta}$ terá distribuição normal multivariada. Por outro lado, fixando um valor para $\mu = \mu_0$ e assumindo que τ^{-2} tem distribuição Gama implicará em uma distribuição t de Student multivariada para $\boldsymbol{\theta}$.

Teoricamente, não há limitação quanto ao número de estágios, mas devido às complexidades resultantes as prioris hierárquicas são especificadas em geral em 2 ou 3 estágios. Além disso, devido à dificuldade de interpretação dos hiperparâmetros em estágios mais altos é prática comum especificar prioris não informativas para este níveis.

Uma aplicação interessante do conceito de hierarquia é quando a informação a priori disponível só pode ser convenientemente resumida através de uma *mistura* de distribuições. Isto implica em considerar uma distribuição discreta para ϕ de modo que, se ϕ assume os possíveis valores ϕ_1, \dots, ϕ_k então

$$p(\theta) = \sum_{i=1}^k p(\theta | \phi_i) p(\phi_i).$$

Não é difícil verificar que a distribuição a posteriori de θ é também uma mistura com veremos a seguir. Aplicando o teorema de Bayes temos que,

$$p(\theta | x) = \frac{p(\theta)p(x|\theta)}{\int p(\theta)p(x|\theta)d\theta} = \frac{\sum_{i=1}^k p(x|\theta)p(\theta|\phi_i)p(\phi_i)}{\sum_{i=1}^k p(\phi_i) \int p(x|\theta)p(\theta|\phi_i)d\theta}.$$

Mas note que a distribuição a posteriori condicional de θ dado ϕ_i é obtida via teorema de Bayes como

$$p(\theta|x, \phi_i) = \frac{p(x|\theta)p(\theta|\phi_i)}{\int p(x|\theta)p(\theta|\phi_i)d\theta} = \frac{p(x|\theta)p(\theta|\phi_i)}{m(x|\phi_i)}$$

e a distribuição a posteriori de ϕ_i é obtida como

$$p(\phi_i) = \frac{m(x|\phi_i)p(\phi)}{p(x)}.$$

Portanto $p(x|\theta)p(\theta|\phi_i)=p(\theta|x, \phi_i)m(x|\phi_i)$. Assim, podemos escrever a posteriori de θ como

$$p(\theta|x) = \frac{\sum_{i=1}^k p(\theta|x, \phi_i)m(x|\phi_i)p(\phi_i)}{\sum_{i=1}^k m(x|\phi_i)p(\phi_i)} = \sum_{i=1}^k p(\theta|x, \phi_i)p(\phi_i|x)$$

Note também que $p(x) = \sum m(x|\phi_i)p(\phi_i)$, isto é a distribuição preditiva, é uma mistura de preditivas condicionais.

Exemplo 2.6 : Se $\theta \in (0, 1)$, a família de distribuições a priori $Beta(a, b)$ é conveniente. Mas estas são sempre unimodais e (se $a \neq b$) assimétricas à esquerda ou à direita. Outras formas interessantes, e mais de acordo com a nossa informação a priori, podem ser obtidas misturando-se 2 ou 3 elementos desta família. Por exemplo,

$$\theta \sim 0,25Beta(3,8) + 0,75Beta(8,3)$$

representa a informação a priori de que $\theta \in (0,5; 0,95)$ com alta probabilidade (0,71) mas também que $\theta \in (0,1; 0,4)$ com probabilidade moderada (0,20). As modas desta distribuição são 0,23 e 0,78. Por outro lado

$$\theta \sim 0,33Beta(4,10) + 0,33Beta(15,28) + 0,33Beta(50,70)$$

representa a informação a priori de que $\theta > 0,6$ com probabilidade desprezível. Estas densidades estão representadas graficamente na Figura 2.2 a seguir. Note que a primeira mistura deu origem a uma distribuição a priori bimodal enquanto a segunda originou uma priori assimétrica à esquerda com média igual a 0,35.

Para outros exemplos de misturas de prioris ver O'Hagan (1994). Para um excelente material sobre modelos hierárquicos ver (Gelman et al. 2004).

2.6 Problemas

1. Mostre que a família de distribuições Beta é conjugada em relação às distribuições amostrais binomial, geométrica e binomial negativa.

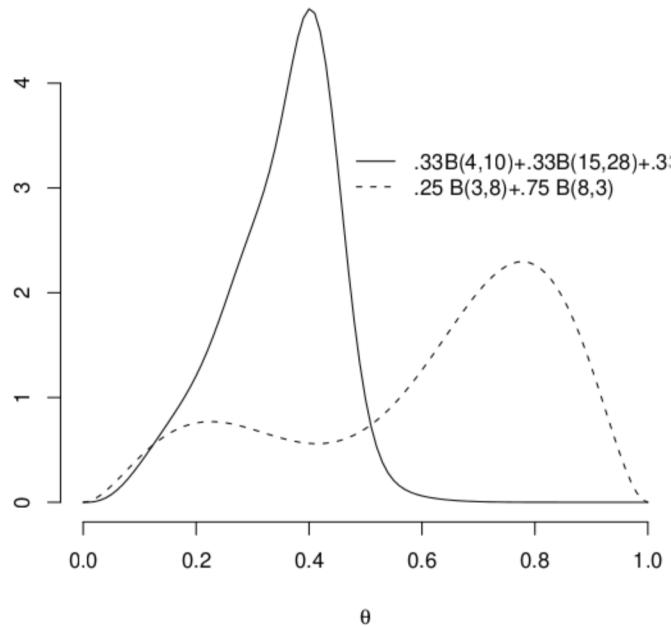


Figura 2.2: Misturas de funções de densidade Beta(3,8) e Beta(8,3) com pesos 0,25 e 0,75 e Beta(4,10), Beta(15,28) e Beta(50,70) com pesos iguais a 0,33.

2. Para uma amostra aleatória de 100 observações da distribuição normal com média θ e desvio-padrão 2 foi especificada uma priori normal para θ .
 - (a) Mostre que o desvio-padrão a posteriori será sempre menor do que $1/5$. Interprete este resultado.
 - (b) Se o desvio-padrão a priori for igual a 1 qual deve ser o menor número de observações para que o desvio-padrão a posteriori seja 0,1?
3. Seja X_1, \dots, X_n uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ conhecido. Utilizando uma distribuição a priori Gama para σ^{-2} com coeficiente de variação 0,5, qual deve ser o tamanho amostral para que o coeficiente de variação a posteriori diminua para 0,1?
4. Seja X_1, \dots, X_n uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ e σ^2 desconhecidos, e considere a priori conjugada de (θ, ϕ) .
 - (a) Determine os parâmetros $(\mu_0, c_0, n_0, \sigma_0^2)$ utilizando as seguintes informações a priori: $E(\theta) = 0$, $P(|\theta| < 1,412) = 0,5$, $E(\phi) = 2$ e $E(\phi^2) = 5$.

- (b) Em uma amostra de tamanho $n = 10$ foi observado $\bar{X} = 1$ e $\sum_{i=1}^n (X_i - \bar{X})^2 = 8$. Obtenha a distribuição a posteriori de θ e esboce os gráficos das distribuições a priori, a posteriori e da função de verossimilhança, com ϕ fixo.
- (c) Calcule $P(|Y| > 1 | \mathbf{x})$ onde Y é uma observação tomada da mesma população.
5. Suponha que o tempo, em minutos, para atendimento a clientes segue uma distribuição exponencial com parâmetro θ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gama com média 0,2 e desvio-padrão 1 para θ .
- Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3,8 minutos, qual a distribuição a posteriori de θ .
 - Qual o menor número de clientes que precisam ser observados para que o coeficiente de variação a posteriori se reduza para 0,1?
6. Seja X_1, \dots, X_n uma amostra aleatória da distribuição de Poisson com parâmetro θ .
- Determine os parâmetros da priori conjugada de θ sabendo que $E(\theta) = 4$ e o coeficiente de variação a priori é 0,5.
 - Quantas observações devem ser tomadas até que a variância a posteriori se reduza para 0,01 ou menos?
 - Mostre que a média a posteriori é da forma $\gamma_n \bar{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \rightarrow 1$ quando $n \rightarrow \infty$. Interprete este resultado.
7. O número médio de defeitos por 100 metros de uma fita magnética é desconhecido e denotado por θ . Atribui-se uma distribuição a priori Gama(2,10) para θ . Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos qual a distribuição a posteriori de θ ?
8. Seja X_1, \dots, X_n uma amostra aleatória da distribuição Bernoulli com parâmetro θ e usamos a priori conjugada $Beta(a, b)$. Mostre que a média a posteriori é da forma $\gamma_n \bar{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \rightarrow 1$ quando $n \rightarrow \infty$. Interprete este resultado.
9. Para uma amostra aleatória X_1, \dots, X_n tomada da distribuição $U(0, \theta)$, mostre que a família de distribuições de Pareto com parâmetros a e b , cuja função de densidade é $p(\theta) = ab^a/\theta^{a+1}$, é conjugada à uniforme.

10. Para uma variável aleatória $\theta > 0$ a família de distribuições Gama-invertida tem função de densidade de probabilidade dada por

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \quad \alpha, \beta > 0.$$

Mostre que esta família é conjugada ao modelo normal com média μ conhecida e variância θ desconhecida.

11. Suponha que $\mathbf{X} = (X_1, X_2, X_3)$ tenha distribuição trinomial com parâmetros n (conhecido) e $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ com $\pi_1 + \pi_2 + \pi_3 = 1$. Mostre que a priori não informativa de Jeffreys para $\boldsymbol{\pi}$ é $p(\boldsymbol{\pi}) \propto [\pi_1 \pi_2 (1 - \pi_1 - \pi_2)]^{-1/2}$.
12. Para cada uma das distribuições abaixo verifique se o modelo é de locação, escala ou locação-escala e obtenha a priori não informativa para os parâmetros desconhecidos.
- (a) Cauchy($0, \beta$).
 - (b) $t_\nu(\mu, \sigma^2)$, ν conhecido.
 - (c) Pareto(a, b), b conhecido.
 - (d) Uniforme $(\theta - 1, \theta + 1)$.
 - (e) Uniforme $(-\theta, \theta)$.
13. Seja uma coleção de variáveis aleatórias independentes X_i com distribuições $p(x_i|\theta_i)$ e seja $p_i(\theta_i)$ a priori não informativa de θ_i , $i = 1, \dots, k$. Mostre que a priori não informativa de Jeffreys para o vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ é dada por $\prod_{i=1}^k p_i(\theta_i)$.
14. Se θ tem priori não informativa $p(\theta) \propto k$, $\theta > 0$ mostre que a priori de $\phi = a\theta + b$, $a \neq 0$ também é $p(\phi) \propto k$.
15. Se θ tem priori não informativa $p(\theta) \propto \theta^{-1}$ mostre que a priori de $\phi = \theta^a$, $a \neq 0$ também é $p(\phi) \propto \phi^{-1}$ e que a priori de $\psi = \log \theta$ é $p(\psi) \propto k$.
16. No Exemplo 1.3, sejam $\phi_i = (\mu_i, \tau_i^2)$, $i = 1, 2$, as médias e variâncias a priori dos físicos A e B respectivamente. As prioris condicionais foram então combinadas como

$$p(\theta) = p(\phi_1)p(\theta|\phi_1) + p(\phi_2)p(\theta|\phi_2)$$

com $p(\phi_1) = 0,25$ e $p(\phi_2) = 0,75$. Usando as posterioris condicionais obtidas naquele exemplo obtenha a distribuição a posteriori de θ (incondicional). Esboce e comente os gráficos das densidades a priori e posteriori.

17. Se $X \sim \text{Binomial Negativa}(v, \theta)$ obtenha a priori de Jeffreys para θ .
18. Se $X \sim \text{Geometrica}(\theta)$ obtenha a priori de Jeffreys para θ .

Capítulo 3

Estimação

A distribuição a posteriori de um parâmetro θ contém toda a informação probabilística a respeito deste parâmetro e um gráfico da sua função de densidade a posteriori é a melhor descrição do processo de inferência. No entanto, algumas vezes é necessário resumir a informação contida na posteriori através de alguns poucos valores numéricos. O caso mais simples é a estimação pontual de θ onde se resume a distribuição a posteriori através de um único número, $\hat{\theta}$. Como veremos a seguir, será mais fácil entender a escolha de $\hat{\theta}$ no contexto de teoria da decisão.

3.1 Introdução à Teoria da Decisão

Um problema de decisão fica completamente especificado pela descrição dos seguintes espaços:

- (i) Espaço do parâmetro ou estados da natureza, Θ .
- (ii) Espaço dos resultados possíveis de um experimento, Ω .
- (iii) Espaço de possíveis ações, A .

Uma regra de decisão δ é uma função definida em Ω que assume valores em A , i.e. $\delta : \Omega \rightarrow A$. A cada decisão δ e a cada possível valor do parâmetro θ podemos associar uma perda $L(\delta, \theta)$ assumindo valores positivos. Definimos assim uma função de perda.

Definição 3.1 *O risco de uma regra de decisão, denotado por $R(\delta)$, é a perda esperada a posteriori, i.e. $R(\delta) = E_{\theta|\mathbf{x}}[L(\delta, \theta)]$.*

Definição 3.2 *Uma regra de decisão δ^* é ótima se tem risco mínimo, i.e. $R(\delta^*) < R(\delta)$, $\forall \delta$. Esta regra será denominada regra de Bayes e seu risco, risco de Bayes.*

Exemplo 3.1: Um laboratório farmacêutico deve decidir pelo lançamento ou não de uma nova droga no mercado. É claro que o laboratório só lançará a droga se achar que ela é eficiente mas isto é exatamente o que é desconhecido. Podemos associar um parâmetro θ aos estados da natureza: droga é eficiente ($\theta = 1$), droga não é eficiente ($\theta = 0$) e as possíveis ações como lança a droga ($\delta = 1$), não lança a droga ($\delta = 0$). Suponha que foi possível construir a seguinte tabela de perdas levando em conta a eficiência da droga,

	eficiente	não eficiente
lança	-500	600
não lança	1500	100

Vale notar que estas perdas traduzem uma avaliação subjetiva em relação à gravidade dos erros cometidos. Suponha agora que a incerteza sobre os estados da natureza é descrita por $P(\theta = 1) = \pi$, $0 < \pi < 1$ avaliada na distribuição atualizada de θ (seja a priori ou a posteriori). Note que, para δ fixo, $L(\delta, \theta)$ é uma variável aleatória discreta assumindo apenas dois valores com probabilidades π e $1 - \pi$. Assim, usando a definição de risco obtemos que

$$\begin{aligned} R(\delta = 0) &= E(L(0, \theta)) = \pi 1500 + (1 - \pi) 100 = 1400\pi + 100 \\ R(\delta = 1) &= E(L(1, \theta)) = \pi(-500) + (1 - \pi) 600 = -1100\pi + 600 \end{aligned}$$

Uma questão que se coloca aqui é, para que valores de π a regra de Bayes será de lançar a droga. Não é difícil verificar que as duas ações levarão ao mesmo risco, i.e. $R(\delta = 0) = R(\delta = 1)$ se somente se $\pi = 0,20$. Além disso, para $\pi < 0,20$ temos que $R(\delta = 0) < R(\delta = 1)$ e a regra de Bayes consiste em não lançar a droga enquanto que $\pi > 0,20$ implica em $R(\delta = 1) < R(\delta = 0)$ e a regra de Bayes deve ser de lançar a droga.

3.2 Estimadores de Bayes

Seja agora uma amostra aleatória X_1, \dots, X_n tomada de uma distribuição com função de (densidade) de probabilidade $p(x|\theta)$ onde o valor do parâmetro θ é desconhecido. Em um problema de inferência como este o valor de θ deve ser estimado a partir dos valores observados na amostra.

Se $\theta \in \Theta$ então é razoável que os possíveis valores de um estimador $\delta(\mathbf{X})$ também devam pertencer ao espaço Θ . Além disso, um bom estimador é aquele para o qual, com alta probabilidade, o erro $\delta(\mathbf{X}) - \theta$ estará próximo de zero. Para cada possível valor de θ e cada possível estimativa $a \in \Theta$ vamos associar uma perda $L(a, \theta)$ de modo que quanto maior a distância entre a e θ maior o

valor da perda. Neste caso, a perda esperada a posteriori é dada por

$$E[L(a, \theta) | \mathbf{x}] = \int L(a, \theta) p(\theta | \mathbf{x}) d\theta$$

e a regra de Bayes consiste em escolher a estimativa que minimiza esta perda esperada.

Aqui vamos discutir apenas funções de perda simétricas, já que estas são mais comumente utilizadas (para outras funções de perda ver por exemplo (Bernardo & Smith 1994) e O'Hagan 1994). Dentre estas a mais utilizada em problemas de estimação é certamente a função de perda quadrática, definida como $L(a, \theta) = (a - \theta)^2$. Neste caso, pode-se mostrar que o estimador de Bayes para o parâmetro θ será a média de sua distribuição atualizada.

Exemplo 3.2: Suponha que queremos estimar a proporção θ de itens defeituosos em um grande lote. Para isto será tomada uma amostra aleatória X_1, \dots, X_n de uma distribuição de Bernoulli com parâmetro θ . Usando uma priori conjugada Beta(α, β) sabemos que após observar a amostra a distribuição a posteriori é Beta($\alpha + t, \beta + n - t$) onde $t = \sum_{i=1}^n x_i$. A média desta distribuição Beta é dada por $(\alpha + t)/(\alpha + \beta + n)$ e portanto o estimador de Bayes de θ usando perda quadrática é

$$\delta(\mathbf{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}.$$

A perda quadrática é às vezes criticada por penalizar demais o erro de estimação. A função de perda absoluta, definida como $L(a, \theta) = |a - \theta|$, introduz punições que crescem linearmente com o erro de estimação e pode-se mostrar que o estimador de Bayes associado é a mediana da distribuição atualizada de θ .

Para reduzir ainda mais o efeito de erros de estimação grandes podemos considerar funções que associam uma perda fixa a um erro cometido, não importando sua magnitude. Uma tal função de perda, denominada perda 0-1, é definida como

$$L(a, \theta) = \begin{cases} 1 & \text{se } |a - \theta| > \epsilon \\ 0 & \text{se } |a - \theta| < \epsilon \end{cases}$$

para todo $\epsilon > 0$. Neste caso pode-se mostrar que o estimador de Bayes é a moda da distribuição atualizada de θ . A moda da posteriori de θ também é chamado de estimador de máxima verossimilhança generalizado (EMVG) e é o mais fácil de ser obtido dentre os estimadores vistos até agora. No caso contínuo devemos obter a solução da equação

$$\frac{\partial p(\theta | \mathbf{x})}{\partial \theta} = 0.$$

Note que isto equivale a obter a solução de

$$\frac{\partial p(\mathbf{x}|\theta)p(\theta)}{\partial\theta}=0$$

e não é necessário conhecer a expressão exata de $p(\theta|\mathbf{x})$.

Exemplo 3.3: Se X_1, \dots, X_n é uma amostra aleatória da $N(\theta, \sigma^2)$ com σ^2 conhecido e usarmos a priori conjugada, i.e. $\theta \sim N(\mu_0, \tau_0^2)$ então a posteriori também será normal e neste caso média, mediana e moda coincidem. Portanto, o estimador de Bayes de θ é dado por

$$\delta(\mathbf{X}) = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\bar{\mathbf{X}}}{\tau_0^{-2} + n\sigma^{-2}}.$$

Exemplo 3.4: No exemplo 3.2 suponha que foram observados 100 itens dos quais 10 eram defeituosos. Usando perda quadrática a estimativa de Bayes de θ é

$$\delta(\mathbf{x}) = \frac{\alpha + 10}{\alpha + \beta + 100}$$

Assim, se a priori for Beta(1,1), ou equivalentemente $U(0, 1)$, então $\delta(\mathbf{x}) = 0,108$. Por outro lado se especificarmos uma priori Beta(1,2), que é bem diferente da anterior, então $\delta(\mathbf{x}) = 0,107$. Ou seja, as estimativas de Bayes são bastante próximas, e isto é uma consequência do tamanho amostral ser grande. Note também que ambas as estimativas são próximas da proporção amostral de defeituosos 0,1, que é a estimativa de máxima verossimilhança. Se usarmos perda 0-1 e priori Beta(1,1) então $\delta(\mathbf{x}) = 0,1$.

3.3 Estimação por Intervalos

Voltamos a enfatizar que a forma mais adequada de expressar a informação que se tem sobre um parâmetro é através de sua distribuição a posteriori. A principal restrição da estimação pontual é que quando estimamos um parâmetro através de um único valor numérico toda a informação presente na distribuição a posteriori é resumida através deste número. É importante também associar alguma informação sobre o quanto precisa é a especificação deste número. Para os estimadores vistos aqui as medidas de incerteza mais usuais são a variância ou o coeficiente de variação para a média a posteriori, a medida de informação observada de Fisher para a moda a posteriori, e a distância entre quartis para a mediana a posteriori.

Nesta seção vamos introduzir um compromisso entre o uso da própria distribuição a posteriori e uma estimativa pontual. Será discutido o conceito de

intervalo de credibilidade (ou intervalo de confiança Bayesiano) baseado no distribuição a posteriori.

Definição 3.3 *C é um intervalo de credibilidade de $100(1-\alpha)\%$, ou nível de credibilidade (ou confiança) $1 - \alpha$, para θ se $P(\theta \in C) \geq 1 - \alpha$.*

Note que a definição expressa de forma probabilística a pertinência ou não de θ ao intervalo. Assim, quanto menor for o tamanho do intervalo mais concentrada é a distribuição do parâmetro, ou seja o tamanho do intervalo informa sobre a dispersão de θ . Além disso, a exigência de que a probabilidade acima possa ser maior do que o nível de confiança é essencialmente técnica pois queremos que o intervalo seja o menor possível, o que em geral implica em usar uma igualdade. No entanto, a desigualdade será útil se θ tiver uma distribuição discreta onde nem sempre é possível satisfazer a igualdade.

Outro fato importante é que os intervalos de credibilidade são invariantes a transformações 1 a 1, $\phi(\theta)$. Ou seja, se $C = [a, b]$ é um intervalo de credibilidade $100(1-\alpha)\%$ para θ então $[\phi(a), \phi(b)]$ é um intervalo de credibilidade $100(1-\alpha)\%$ para $\phi(\theta)$. Note que esta propriedade também vale para intervalos de confiança na inferência clássica.

É possível construir uma infinidade de intervalos usando a definição acima mas estamos interessados apenas naquele com o menor comprimento possível. Pode-se mostrar que intervalos de comprimento mínimo são obtidos tomando-se os valores de θ com maior densidade a posteriori, e esta idéia é expressa matematicamente na definição abaixo.

Definição 3.4 *Um intervalo de credibilidade C de $100(1-\alpha)\%$ para θ é de máxima densidade a posteriori (MDP) se $C = \{\theta \in \Theta : p(\theta|\mathbf{x}) \geq k(\alpha)\}$ onde $k(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$.*

Usando esta definição, todos os pontos dentro do intervalo MDP terão densidade maior do que qualquer ponto fora do intervalo. Além disso, no caso de distribuições com duas caudas, e.g. normal, t de Student, o intervalo MDP é obtido de modo que as caudas tenham a mesma probabilidade. Um problema com os intervalos MDP é que eles não são invariantes a transformações 1 a 1, a não ser para transformações lineares. O mesmo problema ocorre com intervalos de comprimento mínimo na inferência clássica.

3.4 Estimação no Modelo Normal

Os resultados desenvolvidos nos capítulos anteriores serão aplicados ao modelo normal para estimação da média e variância em problemas de uma ou mais

amostras e em modelos de regressão linear. A análise será feita com priori conjugada e priori não informativa quando serão apontadas as semelhanças com a análise clássica. Assim como nos capítulos anteriores a abordagem aqui é introdutória. Um tratamento mais completo do enfoque Bayesiano em modelos lineares pode ser encontrado em Broemeling (1985) e Box & Tiao (1992).

Nesta seção considere uma amostra aleatória X_1, \dots, X_n tomada da distribuição $N(\theta, \sigma^2)$.

3.4.1 Variância Conhecida

Se σ^2 é conhecido e a priori de θ é $N(\mu_0, \tau_0^2)$ então, pelo Teorema 1.1, a posteriori de θ é $N(\mu_1, \tau_1^2)$. Intervalos de confiança Bayesianos para θ podem então ser construídos usando o fato de que

$$\frac{\theta - \mu_1}{\tau_1} | \mathbf{x} \sim N(0, 1).$$

Assim, usando uma tabela da distribuição normal padronizada podemos obter o valor do percentil $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} \leq \frac{\theta - \mu_1}{\tau_1} \leq z_{\alpha/2}\right) = 1 - \alpha$$

e após isolar θ , obtemos que

$$P(\mu_1 - z_{\alpha/2}\tau_1 \leq \theta \leq \mu_1 + z_{\alpha/2}\tau_1) = 1 - \alpha.$$

Portanto $(\mu_1 - z_{\alpha/2}\tau_1; \mu_1 + z_{\alpha/2}\tau_1)$ é o intervalo de confiança $100(1-\alpha)\%$ MDP para θ , devido à simetria da normal.

A priori não informativa pode ser obtida fazendo-se a variância da priori tender a infinito, i.e. $\tau_0^2 \rightarrow \infty$. Neste caso, é fácil verificar que $\tau_1^{-2} \rightarrow n\sigma^{-2}$ e $\mu_1 \rightarrow \bar{x}$, i.e. a média e a precisão da posteriori convergem para a média e a precisão amostrais. Média, moda e mediana a posteriori coincidem então com a estimativa clássica de máxima verossimilhança, \bar{x} . O intervalo de confiança Bayesiano $100(1-\alpha)\%$ é dado por

$$(\bar{x} - z_{\alpha/2} \sigma / \sqrt{n}; \bar{x} + z_{\alpha/2} \sigma / \sqrt{n})$$

e também coincide numericamente com o intervalo de confiança clássico. Aqui entretanto a interpretação do intervalo é como uma afirmação probabilística sobre θ .

3.4.2 Média e Variância desconhecidas

Neste caso, usando a priori conjugada Normal-Gama vista no Capítulo 2 temos que a distribuição a posteriori marginal de θ é dada por

$$\theta|\mathbf{x} \sim t_{n_1}(\mu_1, \sigma_1^2/c_1).$$

Portanto, média, moda e mediana a posteriori coincidem e são dadas por μ_1 . Denotando por $t_{\alpha/2,n_1}$ o percentil $100(1-\alpha/2)\%$ da distribuição $t_{n_1}(0, 1)$ podemos obter este percentil tal que

$$P\left(-t_{\alpha/2,n_1} \leq \sqrt{c_1} \frac{\theta - \mu_1}{\sigma_1} \leq t_{\alpha/2,n_1}\right) = 1 - \alpha$$

e após isolar θ , usando a simetria da distribuição t -Student obtemos que

$$\left(\mu_1 - t_{\alpha/2,n_1} \frac{\sigma_1}{\sqrt{c_1}} \leq \theta \leq \mu_1 + t_{\alpha/2,n_1} \frac{\sigma_1}{\sqrt{c_1}}\right)$$

é o intervalo de confiança Bayesiano $100(1-\alpha)\%$ de MDP para θ .

No caso da variância populacional σ^2 intervalos de confiança podem ser obtidos usando os percentis da distribuição qui-quadrado uma vez que a distribuição a posteriori de ϕ é tal que $n_1\sigma_1^2\phi|\mathbf{x} \sim \chi_{n_1}^2$. Denotando por

$$\underline{\chi}_{\alpha/2,n_1}^2 \quad \text{e} \quad \bar{\chi}_{\alpha/2,n_1}^2$$

os percentis $\alpha/2$ e $1-\alpha/2$ da distribuição qui-quadrado com n_1 graus de liberdade respectivamente, podemos obter estes percentis tais que

$$P\left(\frac{\underline{\chi}_{\alpha/2,n_1}^2}{n_1\sigma_1^2} \leq \phi \leq \frac{\bar{\chi}_{\alpha/2,n_1}^2}{n_1\sigma_1^2}\right) = 1 - \alpha.$$

Note que este intervalo não é de MDP já que a distribuição qui-quadrado não é simétrica. Como $\sigma^2 = 1/\phi$ é uma função 1 a 1 podemos usar a propriedade de invariância e portanto

$$\left(\frac{n_1\sigma_1^2}{\bar{\chi}_{\alpha/2,n_1}^2}; \frac{n_1\sigma_1^2}{\underline{\chi}_{\alpha/2,n_1}^2}\right)$$

é o intervalo de confiança Bayesiano $100(1-\alpha)\%$ para σ^2 .

Um caso particular é quanto utilizamos uma priori não informativa. Vimos na Seção 2.4 que a priori não informativa de locação e escala é $p(\theta, \sigma) \propto 1/\sigma$, portanto pela propriedade de invariância segue que a priori não informativa de (θ, ϕ) é obtida fazendo-se $p(\theta, \phi) \propto \phi^{-1}$ pois $p(\theta, \sigma^2) \propto \sigma^{-2}$. Note que este é um caso particular (degenerado) da priori conjugada natural com $c_0 = 0$, $\sigma_0^2 = 0$ e

$n_0 = -1$. Neste caso a distribuição a posteriori marginal de θ fica

$$\theta|\mathbf{x} \sim t_{n-1}(\bar{x}, s^2/n)$$

sendo $s^2 = 1/(n-1) \sum_{i=1}^n (x_i - \bar{x})^2$ a variância amostral. Mais uma vez média, moda e mediana a posteriori de θ coincidem com a média amostral \bar{x} que é a estimativa de máxima verossimilhança. Como $\sqrt{n}(\theta - \bar{x})/s \sim t_{n-1}(0, 1)$ segue que o intervalo de confiança 100(1- α)% para θ de MDP é

$$\left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

que coincide numericamente com o intervalo de confiança clássico.

Para fazer inferências sobre σ^2 temos que

$$\phi|\mathbf{x} \sim \text{Gama} \left(\frac{n-1}{2}, \frac{(n-1)s^2}{2} \right) \quad \text{ou} \quad (n-1)s^2\phi|\mathbf{x} \sim \chi^2_{n-1}.$$

A estimativa pontual de σ^2 utilizada é $[E(\phi|x)]^{-1} = s^2$ que coincide com a estimativa clássica uma vez que o estimador de máxima verossimilhança $(n-1)S^2/n$ é viciado e normalmente substituído por S^2 (que é não viciado). Os intervalos de confiança 100(1- α)% Bayesiano e clássico também coincidem e são dados por

$$\left(\frac{(n-1)s^2}{\bar{\chi}_{\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\underline{\chi}_{\alpha/2, n-1}^2} \right).$$

Mais uma vez vale enfatizar que esta coincidência com as estimativas clássicas é apenas numérica uma vez que as interpretações dos intervalos diferem radicalmente.

3.4.3 O Caso de duas Amostras

Nesta seção vamos assumir que X_{11}, \dots, X_{1n_1} e X_{21}, \dots, X_{2n_2} são amostras aleatórias das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ respectivamente e que as amostras são independentes.

Para começar vamos assumir que as variâncias σ_1^2 e σ_2^2 são conhecidas. Neste caso, a função de verossimilhança é dada por

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2 | \theta_1, \theta_2) &= p(\mathbf{x}_1 | \theta_1)p(\mathbf{x}_2 | \theta_2) \\ &\propto \exp \left\{ -\frac{n_1}{2\sigma_1^2}(\theta_1 - \bar{x}_1)^2 \right\} \exp \left\{ -\frac{n_2}{2\sigma_2^2}(\theta_2 - \bar{x}_2)^2 \right\} \end{aligned}$$

isto é, o produto de verossimilhanças relativas a θ_1 e θ_2 . Assim, se assumirmos que θ_1 e θ_2 são independentes a priori então eles também serão independentes a

posteriori já que

$$p(\theta_1, \theta_2 | \mathbf{x}_1, \mathbf{x}_2) = \frac{p(\mathbf{x}_1 | \theta_1)p(\theta_1)}{p(\mathbf{x}_1)} \times \frac{p(\mathbf{x}_2 | \theta_2)p(\theta_2)}{p(\mathbf{x}_2)}.$$

Se usarmos a classe de prioris conjugadas $\theta_i \sim N(\mu_i, \tau_i^2)$ então as posterioris independentes serão $\theta_i | \mathbf{x}_i \sim N(\mu_i^*, \tau_i^{*2})$ onde

$$\mu_i^* = \frac{\tau_i^{-2}\mu_i + n_i\sigma_i^{-2} \bar{\mathbf{x}}_i}{\tau_i^{-2} + n_i\sigma_i^{-2}} \quad \text{e} \quad \tau_i^{*2} = 1/(\tau_i^{-2} + n_i\sigma_i^{-2}), \quad i = 1, 2.$$

Em geral estaremos interessados em comparar as médias populacionais, i.e queremos estimar $\beta = \theta_1 - \theta_2$ (por exemplo, testar se $\theta_1 = \theta_2$). Neste caso, a posteriori de β é facilmente obtida, devido à independência, como

$$\beta | \mathbf{x}_1, \mathbf{x}_2 \sim N(\mu_1^* - \mu_2^*, \tau_1^{*2} + \tau_2^{*2})$$

e podemos usar $\mu_1^* - \mu_2^*$ como estimativa pontual para a diferença e também construir um intervalo de credibilidade MDP para esta diferença.

$$(\mu_1^* - \mu_2^*) \pm z_{\alpha/2} \sqrt{\tau_1^{*2} + \tau_2^{*2}}.$$

Note que se usarmos priori não informativa, i.e. fazendo $\tau_i^2 \rightarrow \infty$, $i = 1, 2$ então a posteriori fica

$$\beta | \mathbf{x}_1, \mathbf{x}_2 \sim N\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

e o intervalo obtido coincidirá mais uma vez com o intervalo de confiança clássico.

No caso de variâncias populacionais desconhecidas porém iguais, temos que $\phi = \sigma_1^{-2} = \sigma_2^{-2} = \sigma^{-2}$. A priori conjugada pode ser construída em duas etapas. No primeiro estágio, assumimos que, dado ϕ , θ_1 e θ_2 são a priori condicionalmente independentes, e especificamos

$$\theta_i | \phi \sim N(\mu_i, (c_i\phi)^{-1}), i = 1, 2.$$

e no segundo estágio, especificamos a priori conjugada natural para ϕ , i.e.

$$\phi \sim \text{Gama}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Combinando as prioris acima não é difícil verificar que a priori conjunta de

$(\theta_1, \theta_2, \phi)$ é

$$\begin{aligned} p(\theta_1, \theta_2, \phi) &= p(\theta_1|\phi)p(\theta_2|\phi)p(\phi) \\ &\propto \phi^{n_0/2} \exp \left\{ -\frac{\phi}{2} \left[n_0 \sigma_0^2 + c_1 (\theta_1 - \mu_1)^2 + c_2 (\theta_2 - \mu_2)^2 \right] \right\}. \end{aligned}$$

Além disso, também não é difícil obter a priori condicional de $\beta = \theta_1 - \theta_2$, dado ϕ , como

$$\beta|\phi \sim N(\mu_1 - \mu_2, \phi^{-1}(c_1^{-1} + c_2^{-1}))$$

e portanto, usando os resultados da Seção 2.3.5 segue que a distribuição a priori marginal da diferença é

$$\beta \sim t_{n_0}(\mu_1 - \mu_2, \sigma_0^2(c_1^{-1} + c_2^{-1})).$$

Podemos mais uma vez obter a posteriori conjunta em duas etapas já que θ_1 e θ_2 também serão condicionalmente independentes a posteriori, dado ϕ . Assim, no primeiro estágio usando os resultados obtidos anteriormente para uma amostra segue que

$$\theta_i|\phi, \mathbf{x} \sim N(\mu_i^*, (c_i^* \phi)^{-1}), \quad i = 1, 2$$

onde

$$\mu_i^* = \frac{c_i \mu_i + n_i \bar{x}_i}{c_i + n_i} \quad \text{e} \quad c_i^* = c_i + n_i.$$

Na segunda etapa temos que combinar a verossimilhança com a priori de $(\theta_1, \theta_2, \phi)$. Definindo a variância amostral combinada

$$s_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

e denotando $\nu = n_1 + n_2 - 2$, a função de verossimilhança pode ser escrita como

$$p(\mathbf{x}_1, \mathbf{x}_2 | \theta_1, \theta_2, \phi) = \phi^{(n_1+n_2)/2} \exp \left\{ -\frac{\phi}{2} \left[\nu s^2 + n_1 (\theta_1 - \bar{x}_1)^2 + n_2 (\theta_2 - \bar{x}_2)^2 \right] \right\}$$

e após algum algebrismo obtemos que a posteriori é proporcional a

$$\phi^{(n_0+n_1+n_2)/2} \exp \left\{ -\frac{\phi}{2} \left[n_0 \sigma_0^2 + \nu s^2 + \sum_{i=1}^2 \frac{c_i n_i}{c_i^*} (\mu_i - \bar{x}_i)^2 + c_i^* (\theta_i - \mu_i^*)^2 \right] \right\}.$$

Como esta posteriori tem o mesmo formato da priori segue por analogia que

$$\phi | \mathbf{x} \sim \text{Gama} \left(\frac{n_0^*}{2}, \frac{n_0^* \sigma_0^{*2}}{2} \right)$$

onde $n_0^* = n_0 + n_1 + n_2$ e $n_0^* \sigma_0^{*2} = n_0 \sigma_0^2 + \nu s^2 + \sum_{i=1}^2 c_i n_i (\mu_i - \bar{\mathbf{x}}_i)^2 / c_i^*$. Ainda por analogia com o caso de uma amostra, a posteriori marginal da diferença é dada por

$$\beta | \mathbf{x} \sim t_{n_0^*}(\mu_1^* - \mu_2^*, \sigma_0^{*2} (c_1^{*-1} + c_2^{*-1})).$$

Assim, média, moda e mediana a posteriori de β coincidem e a estimativa pontual é $\mu_1^* - \mu_2^*$. Também intervalos de credibilidade de MDP podem ser obtidos usando os percentis da distribuição t de Student. Para a variância populacional a estimativa pontual usual é σ_0^{*2} e intervalos podem ser construídos usando os percentis da distribuição qui-quadrado já que $n_0^* \sigma_0^{*2} \phi | \mathbf{x} \sim \chi_{n_0^*}^2$.

Vejamos agora como fica a análise usando priori não informativa. Neste caso, $p(\theta_1, \theta_2, \phi) \propto \phi^{-1}$ e isto equivale a um caso particular (degenerado) da priori conjugada com $c_i = 0$, $\sigma_0^2 = 0$ e $n_0 = -2$. Assim, temos que $c_i^* = n_i$, $\mu_i^* = \bar{\mathbf{x}}_i$, $n_0^* = \nu$ e $n_0^* \sigma_0^{*2} = \nu s^2$ e a estimativa pontual concide com a estimativa de máxima verossimilhança $\hat{\beta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. O intervalo de $100(1 - \alpha)\%$ de MDP para β tem limites

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 \pm t_{\frac{\alpha}{2}, \nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

que coincide numericamente com o intervalo de confiança clássico.

O intervalo de $100(1 - \alpha)\%$ para σ^2 é obtido de maneira análoga ao caso de uma amostra usando a distribuição qui-quadrado, agora com ν graus de liberdade, i.e.

$$\left(\frac{\nu s_p^2}{\bar{\chi}_{\frac{\alpha}{2}, \nu}^2}, \frac{\nu s_p^2}{\underline{\chi}_{\frac{\alpha}{2}, \nu}^2} \right).$$

3.4.4 Variâncias desiguais

Até agora assumimos que as variâncias populacionais desconhecidas eram iguais (ou pelo menos aproximadamente iguais). Na inferência clássica a violação desta suposição leva a problemas teóricos e práticos uma vez que não é trivial encontrar uma quantidade pivotal para β com distribuição conhecida ou tabelada. Na verdade, se existem grandes diferenças de variabilidade entre as duas populações pode ser mais apropriado analisar conjuntamente as consequências das diferenças entre as médias e as variâncias. Assim, caso o pesquisador tenha interesse no parâmetro β deve levar em conta os problemas de ordem teórica introduzidos por uma diferença substancial entre σ_1^2 e σ_2^2 .

Do ponto de vista Bayesiano o que precisamos fazer é combinar informação a priori com a verossimilhança e basear a estimação na distribuição a posteriori. A função de verossimilhança agora pode ser fatorada como

$$p(\mathbf{x}_1, \mathbf{x}_2 | \theta_1, \theta_2, \sigma_1^2, \sigma_2^2) = p(\mathbf{x}_1 | \theta_1, \sigma_1^2) p(\mathbf{x}_2 | \theta_2, \sigma_2^2)$$

e vamos adotar prioris conjugadas normal-gama independentes com parâmetros $(\mu_i, c_i, \nu_i, \sigma_{0i}^2)$ para cada uma das amostras. Fazendo as operações usuais para cada amostra, e usando a conjugação da normal-gama, obtemos as seguintes distribuições a posteriori independentes

$$\theta_i | \mathbf{x} \sim t_{n_{0i}^*}(\mu_i^*, \sigma_{0i}^{*2}/c_i^*) \quad \text{e} \quad \phi_i | \mathbf{x} \sim \text{Gama}\left(\frac{n_{0i}^*}{2}, \frac{n_{0i}^* \sigma_{0i}^{*2}}{2}\right), \quad i = 1, 2.$$

Pode-se mostrar que β tem uma distribuição a posteriori chamada Behrens-Fisher, que é semelhante à t de Student e é tabelada. Assim, intervalos de credibilidade podem ser construídos usando-se estes valores tabelados.

Outra situação de interesse é a comparação das duas variâncias populacionais. Neste caso, faz mais sentido utilizar a razão de variâncias ao invés da diferença já que elas medem a escala de uma distribuição e são sempre positivas. Neste caso temos que obter a distribuição a posteriori de $\sigma_2^2/\sigma_1^2 = \phi_1/\phi_2$. Usando a independência a posteriori de ϕ_1 e ϕ_2 e após algum algebrismo pode-se mostrar que

$$\frac{\sigma_{01}^{*2} \phi_1}{\sigma_{02}^{*2} \phi_2} \sim F(n_{01}^*, n_{02}^*).$$

Embora sua função de distribuição não possa ser obtida analiticamente os valores estão tabelados em muitos livros de estatística e também podem ser obtidos na maioria dos pacotes computacionais. Os percentis podem então ser utilizados na construção de intervalos de credibilidade para a razão de variâncias.

Uma propriedade bastante útil para calcular probabilidade com a distribuição F vem do fato de que se $X \sim F(\nu_2, \nu_1)$ então $X^{-1} \sim F(\nu_1, \nu_2)$ por simples inversão na razão de distribuições qui-quadrado independentes. Assim, denotando os quantis α e $1 - \alpha$ da distribuição $F(\nu_1, \nu_2)$ por $F_\alpha(\nu_1, \nu_2)$ e $\bar{F}_\alpha(\nu_1, \nu_2)$ respectivamente segue que

$$F_\alpha(\nu_1, \nu_2) = \frac{1}{\bar{F}_\alpha(\nu_2, \nu_1)}.$$

Note que é usual que os livros forneçam tabelas com os percentis superiores da distribuição F para várias combinações de valores de ν_1 e ν_2 devido à propriedade acima. Por exemplo, se temos os valores tabelados dos quantis 0,95 podemos obter também um quantil 0,05. Basta procurar o quantil 0,95 inverterndo os graus de liberdade.

Finalmente, a análise usando priori não informativa pode ser feita para $p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2} \sigma_2^{-2}$ e será deixada como exercício.

3.5 Exercícios

1. Gere 2 amostras de tamanho 50 da distribuição $N(0, 1)$. Agora construa um intervalo MDP de 95% para a diferença entre as médias (assuma variância conhecida igual a 1). Qual a sua conclusão?
2. Repita a análise da Seção 3.4.4 usando priori não informativa para $p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2} \sigma_2^{-2}$.

Capítulo 4

Métodos Aproximados

4.1 Computação Bayesiana

Existem várias formas de resumir a informação descrita na distribuição a posteriori. Esta etapa frequentemente envolve a avaliação de probabilidades ou esperanças.

Neste capítulo serão descritos métodos baseados em simulação, incluindo Monte Carlo simples, Monte Carlo com função de importância, métodos de reamostragem e Monte Carlo via cadeias de Markov (MCMC). O material apresentado é introdutório e mais detalhes sobre os estes métodos podem ser obtidos por exemplo em Gamerman (1997), Robert & Casella (1999) e Gamerman & Lopes (2006). Outros métodos computacionalmente intensivos como técnicas de otimização e integração numérica, bem como aproximações analíticas não serão tratados aqui e uma referência introdutória é Migon & Gamerman (1999).

Todos os algoritmos que serão vistos aqui são não determinísticos, i.e. todos requerem a simulação de números (pseudo) aleatórios de alguma distribuição de probabilidades. Em geral, a única limitação para o número de simulações são o tempo de computação e a capacidade de armazenamento dos valores simulados. Assim, se houver qualquer suspeita de que o número de simulações é insuficiente, a abordagem mais simples consiste em simular mais valores.

4.2 Uma Palavra de Cautela

Apesar da sua grande utilidade, os métodos que serão apresentados aqui devem ser aplicados com cautela. Devido à facilidade com que os recursos computacionais podem ser utilizados hoje em dia, corremos o risco de apresentar uma solução para o problema errado (o erro tipo 3) ou uma solução ruim para o problema certo. Assim, os métodos computacionalmente intensivos não devem ser vistos como substitutos do pensamento crítico sobre o problema por parte do pesquisador.

Além disso, sempre que possível deve-se utilizar soluções exatas, i.e. não aproximadas, se elas existirem. Por exemplo, em muitas situações em que precisamos calcular uma integral múltipla existe solução exata em algumas dimensões, enquanto nas outras dimensões temos que usar métodos de aproximação.

4.3 O Problema Geral da Inferência Bayesiana

A distribuição a posteriori pode ser convenientemente resumida em termos de esperanças de funções particulares do parâmetro θ , i.e.

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

ou distribuições a posteriori marginais quando $\boldsymbol{\theta}$ for multidimensional, por exemplo se $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ então

$$p(\boldsymbol{\theta}_1|\mathbf{x}) = \int p(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}_2.$$

Assim, o problema geral da inferência Bayesiana consiste em calcular tais valores esperados segundo a distribuição a posteriori de θ . Alguns exemplos são,

1. Constante normalizadora. $g(\theta) = 1$ e $p(\theta|\mathbf{x}) = kq(\theta)$, segue que

$$k = \left[\int q(\theta)d\theta \right]^{-1}.$$

2. Se $g(\theta) = \theta$, então têm-se $\mu = E(\theta|\mathbf{x})$, média a posteriori.
3. Quando $g(\theta) = (\theta - \mu)^2$, então $\sigma^2 = E((\theta - \mu)^2|\mathbf{x})$, a variância a posteriori.
4. Se $g(\theta) = I_A(\theta)$, onde $I_A(x) = 1$ se $x \in A$ e zero caso contrário, então

$$P(A | \mathbf{x}) = \int_A p(\theta|\mathbf{x})d\theta$$
5. Seja $g(\theta) = p(y|\theta)$, onde $y \perp \mathbf{x} | \theta$. Nestas condições obtemos $E[p(y|\mathbf{x})]$, a distribuição preditiva de y , uma observação futura.

Portanto, a habilidade de integrar funções, muitas vezes complexas e multidimensionais, é extremamente importante em inferência Bayesiana. Inferência exata somente será possível se estas integrais puderem ser calculadas analiticamente, caso contrário devemos usar aproximações. Nas próximas seções iremos apresentar métodos aproximados baseados em simulação para obtenção dessas integrais.

4.4 Método de Monte Carlo Simples

A idéia do método é justamente escrever a integral que se deseja calcular como um valor esperado. Para introduzir o método considere o problema de calcular a integral de uma função $g(\theta)$ no intervalo (a, b) , i.e.

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

Esta integral pode ser reescrita como

$$I = \int_a^b (b-a)g(\theta) \frac{1}{b-a} d\theta = (b-a)E[g(\theta)]$$

identificando θ como uma variável aleatória com distribuição $U(a, b)$. Assim, transformamos o problema de avaliar a integral no problema estatístico de estimar uma média, $E[g(\theta)]$. Se dispomos de uma amostra aleatória de tamanho n , $\theta_1, \dots, \theta_n$ da distribuição uniforme no intervalo (a, b) teremos também uma amostra de valores $g(\theta_1), \dots, g(\theta_n)$ da função $g(\theta)$ e a integral acima pode ser estimada pela média amostral, i.e.

$$\hat{I} = (b-a) \frac{1}{n} \sum_{i=1}^n g(\theta_i).$$

Não é difícil verificar que esta estimativa é não viésada já que

$$E(\hat{I}) = \frac{(b-a)}{n} \sum_{i=1}^n E[g(\theta_i)] = (b-a)E[g(\theta)] = \int_a^b g(\theta) d\theta.$$

Podemos então usar o seguinte algoritmo

1. gere $\theta_1, \dots, \theta_n$ da distribuição $U(a, b)$;
2. calcule $g(\theta_1), \dots, g(\theta_n)$;
3. calcule a média amostral $\bar{g} = \sum_{i=1}^n g(\theta_i)/n$
4. calcule $\hat{I} = (b-a)\bar{g}$

Exemplo 4.1: Suponha que queremos calcular $\int_1^3 \exp(-x) dx$. A integral pode ser reescrita como

$$(3-1) \int_1^3 \exp(-x)/(3-1) dx$$

e será aproximada usando 100 valores simulados da distribuição Uniforme no intervalo $(1,3)$ e calculando $y_i = e^{-x_i}$, $i = 1, \dots, 100$. O valor aproximado da

integral é $2 \sum_{i=1}^{100} y_i / 100$. Por outro lado, sabemos que $\exp(-x)$ é a função de densidade de uma v.a. $X \sim Exp(1)$ e portanto a integral pode ser calculada de forma exata,

$$\int_1^3 \exp(-x)dx = Pr(X < 3) - Pr(X < 1) = 0.3181.$$

Podemos escrever uma função mais geral no R cujos argumentos são o número de simulações e os limites de integração.

```
> int.exp = function(n, a, b) {
+   x = runif(n, a, b)
+   y = exp(-x)
+   int.exp = (b - a) * mean(y)
+   return(int.exp)
+ }
```

Executando a função `int.exp` digamos 50 vezes com $n = 10$, $a = 1$ e $b = 3$ existirá uma variação considerável na estimativa da integral. Veja a Figura 4.1. Isto se chama “erro de Monte Carlo” e decresce conforme aumentamos o número de simulações. Repetindo o experimento com $n = 1000$ a variação ficará bem menor. Na Figura 4.2 a evolução deste erro conforme se aumenta o número de simulações fica bem evidente. Os comandos do R a seguir foram utilizados.

```
> n = c(20, 50, 100, 200, 500)
> y = matrix(0, ncol = length(n), nrow = 50)
> for (j in 1:length(n)) {
+   m = NULL
+   for (i in 1:50) m = c(m, int.exp(n[j], 1, 3))
+   y[, j] = m
+ }
> boxplot(data.frame(y), names = n)
```

A generalização é bem simples para o caso em que a integral é a esperança matemática de uma função $g(\theta)$ onde θ tem função de densidade $p(\theta)$, i.e.

$$I = \int_a^b g(\theta)p(\theta)d\theta = E[g(\theta)]. \quad (4.1)$$

Neste caso, podemos usar o mesmo algoritmo descrito acima modificando o passo 1 para gerar $\theta_1, \dots, \theta_n$ da distribuição $p(\theta)$ e calculando

$$\hat{I} = \bar{g} = \frac{1}{n} \sum_{i=1}^n g(\theta_i).$$

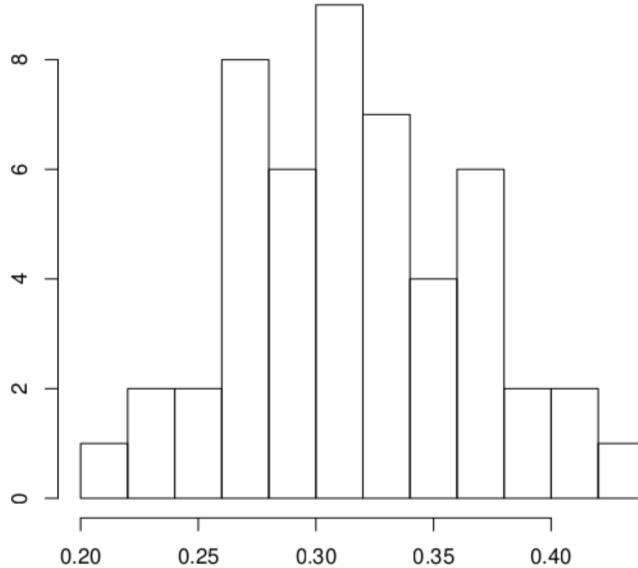


Figura 4.1: Histograma de 50 estimativas de Monte Carlo da integral no Exemplo 4.1 com $n = 10$.

Uma vez que as gerações são independentes, pela Lei Forte dos Grandes Números segue que \hat{I} converge quase certamente para I ,

$$\frac{1}{n} \sum_{i=1}^n g(\theta_i) \rightarrow E[g(\theta)], \quad n \rightarrow \infty.$$

Além disso, temos uma amostra $g(\theta_1), \dots, g(\theta_n)$ tal que

$$E[g(\theta_i)] = E[g(\theta)] = I \quad \text{e} \quad \text{Var}[g(\theta_i)] = \sigma^2 = \frac{1}{n} \sum (g(\theta_i) - \bar{g})^2$$

e portanto a variância do estimador pode também ser estimada como

$$v = \frac{1}{n^2} \sum_{i=1}^n (g(\theta_i) - \bar{g})^2,$$

i.e. a aproximação pode ser tão acurada quanto se deseje bastando aumentar o valor de n . É importante notar que n está sob nosso controle aqui, e não se trata do tamanho da amostra de dados.

O Teorema Central do Limite também se aplica aqui de modo que para n

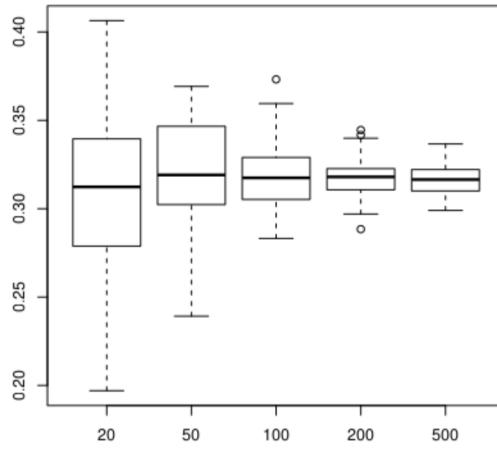


Figura 4.2: Boxplots para 50 estimativas da integral no Exemplo 4.1 com $n=20, 50, 100, 200$, e 500 simulações.

grande segue que

$$\frac{\bar{g} - E[g(\theta)]}{\sqrt{v}}$$

tem distribuição aproximadamente $N(0, 1)$. Podemos usar este resultado para testar convergência e construir intervalos de confiança.

No caso multivariado a extensão também é direta. Seja $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ um vetor aleatório de dimensão k com função de densidade $p(\boldsymbol{\theta})$. Neste caso os valores gerados serão também vetores $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ e o estimador de Monte Carlo fica

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n g(\boldsymbol{\theta}_i)$$

Exemplo 4.2: Suponha que queremos calcular $Pr(X < 1, Y < 1)$ onde o vetor aleatório (X, Y) tem distribuição Normal padrão bivariada com correlação igual a 0,5. Note que esta probabilidade é a integral de $p(x, y)$ definida no intervalo acima, portanto simulando valores desta distribuição poderemos estimar esta probabilidade como a proporção de pontos que caem neste intervalo. A Figura 4.3 apresenta um diagrama de dispersão dos valores simulados e foi obtida usando os comandos do R abaixo.

```

> Sigma = matrix(c(1, 0.5, 0.5, 1), 2, 2)
> m = c(0, 0)
> require(MASS)
> y = mvrnorm(n = 1000, mu = m, Sigma = Sigma)
> plot(y[, 1], y[, 2], xlab = "x", ylab = "y")
> abline(1, 0)
> abline(v = 1)

```

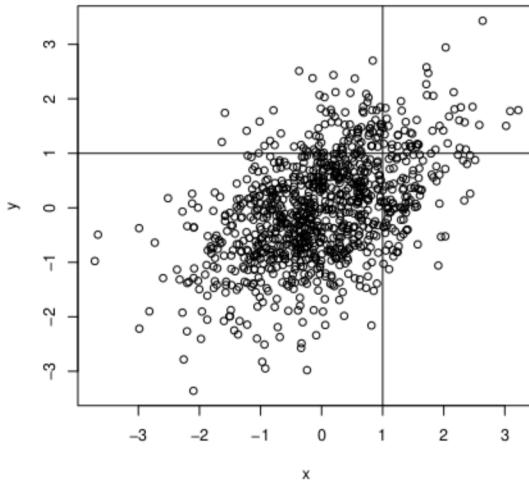


Figura 4.3: Diagrama de dispersão de 1000 valores simulados da distribuição $N(0,1)$ bivariada.

Uma grande vantagem dos métodos de simulação é que após uma amostra de vetores aleatórios ser gerada podemos facilmente calcular características das distribuições marginais e condicionais. No Exemplo 4.2, para calcular $Pr(X < 1)$ basta calcular a frequência relativa de pontos (x_i, y_i) tais que $x_i < 1$. Para calcular a probabilidade condicional $Pr(X < 1|Y < 1)$ basta selecionar somente aqueles pontos cuja segunda coordenada é menor do que 1. Depois calcula-se a frequência relativa dos pontos restantes cuja primeira coordenada é menor do que 1.

4.4.1 Monte Carlo via Função de Importância

Em muitas situações pode ser muito custoso ou mesmo impossível simular valores da distribuição a posteriori. Neste caso, pode-se recorrer à uma função $q(\theta)$ que seja de fácil amostragem, usualmente chamada de *função de importância*. O procedimento é comumente chamado de *amostragem por importância*.

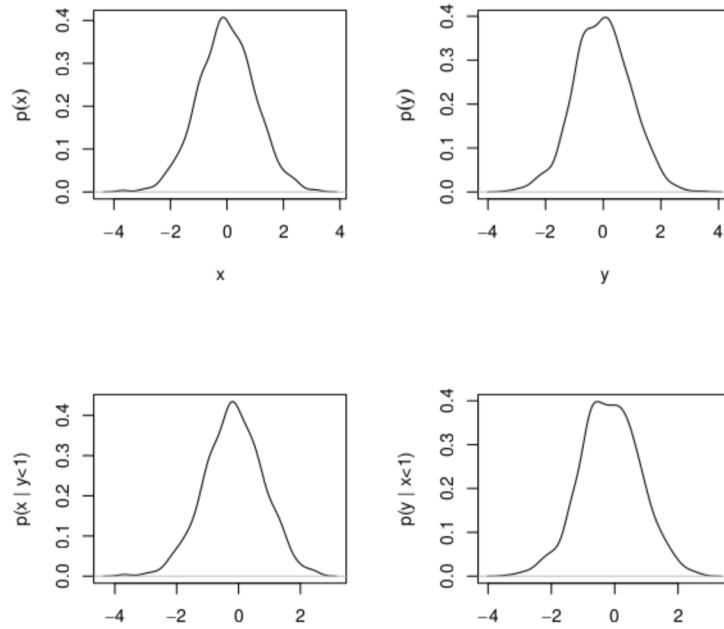


Figura 4.4: Estimativas das densidades marginais e condicionais no Exemplo 4.2.

Se $q(\theta)$ for uma função de densidade definida no mesmo espaço variação de θ então a integral (4.1) pode ser reescrita como

$$I = \int \frac{g(\theta)p(\theta)}{q(\theta)} q(\theta) dx = E \left[\frac{g(\theta)p(\theta)}{q(\theta)} \right]$$

onde a esperança agora é com respeito a distribuição q . Assim, se dispomos de uma amostra aleatória $\theta_1, \dots, \theta_n$ tomada da distribuição q o estimador de Monte Carlo da integral acima fica

$$\hat{I} = \frac{1}{n} \sum_{i=1}^n \frac{g(\theta_i)p(\theta_i)}{q(\theta_i)}.$$

e tem as mesmas propriedades do estimador de Monte Carlo simples.

Em princípio não há restrições quanto à escolha da densidade de importância q , porém na prática alguns cuidados devem ser tomados. Pode-se mostrar que a escolha ótima no sentido de minimizar a variância do estimador consiste em tomar $q(\theta) \propto g(\theta)p(\theta)$.

Exemplo 4.3: Para uma única observação X suponha que

$$X|\theta \sim N(\theta, 1) \quad \text{e} \quad \theta \sim \text{Cauchy}(0, 1).$$

Então,

$$p(x|\theta) \propto \exp[-(x-\theta)^2/2] \quad \text{e} \quad p(\theta) = \frac{1}{\pi(1+\theta^2)}.$$

Portanto, a densidade a posteriori de θ é dada por

$$p(\theta|x) = \frac{\frac{1}{1+\theta^2} \exp[-(x-\theta)^2/2]}{\int \frac{1}{1+\theta^2} \exp[-(x-\theta)^2/2] d\theta}.$$

Suponha agora que queremos estimar θ usando função de perda quadrática. Como vimos no Capítulo 3 isto implica em tomar a média a posteriori de θ como estimativa. Mas

$$E[\theta|\mathbf{x}] = \int \theta p(\theta|\mathbf{x}) d\theta = \frac{\int \frac{\theta}{1+\theta^2} \exp[-(x-\theta)^2/2] d\theta}{\int \frac{1}{1+\theta^2} \exp[-(x-\theta)^2/2] d\theta}$$

e as integrais no numerador e denominador não têm solução analítica exata. Uma solução aproximada via simulação de Monte Carlo pode ser obtida usando o seguinte algoritmo,

1. gerar $\theta_1, \dots, \theta_n$ independentes da distribuição $N(x, 1)$;
2. calcular $g_i = \frac{\theta_i}{1+\theta_i^2}$ e $g_i^* = \frac{1}{1+\theta_i^2}$;
3. calcular $\hat{E}(\theta|\mathbf{x}) = \frac{\sum_{i=1}^n g_i}{\sum_{i=1}^n g_i^*}$.

Este exemplo ilustrou um problema que geralmente ocorre em aplicações Bayesianas. Como a posteriori só é conhecida a menos de uma constante de proporcionalidade as esperanças a posteriori são na verdade uma razão de integrais. Neste caso, a aproximação é baseada na razão dos dois estimadores de Monte Carlo para o numerador e denominador.

Exercícios

1. Para cada uma das distribuições $N(0, 1)$, Gama(2,5) e Beta(2,5) gere 100, 1000 e 5000 valores independentes. Faça um gráfico com o histograma e

a função de densidade superposta. Estime a média e a variância da distribuição. Estime a variância do estimador da média.

2. Para uma única observação X com distribuição $N(\theta, 1)$, θ desconhecido, queremos fazer inferência sobre θ usando uma priori Cauchy(0,1). Gere um valor de X para $\theta = 2$, i.e. $x \sim N(2, 1)$.
 - (a) Estime θ através da sua média a posteriori usando o algoritmo do Exemplo 4.3.
 - (b) Estime a variância da posteriori.
 - (c) Generalize o algoritmo para k observações X_1, \dots, X_k da distribuição $N(\theta, 1)$.

4.5 Métodos de Reamostragem

Existem distribuições para as quais é muito difícil ou mesmo impossível simular valores. A idéia dos métodos de reamostragem é gerar valores em duas etapas. Na primeira etapa gera-se valores de uma distribuição auxiliar conhecida. Na segunda etapa utiliza-se um mecanismo de correção para que os valores sejam representativos (ao menos aproximadamente) da distribuição a posteriori. Na prática costuma-se tomar a priori como distribuição auxiliar conforme proposto em Smith & Gelfand (1992).

4.5.1 Método de Rejeição

Considere uma função de densidade auxiliar $q(\theta)$ da qual sabemos gerar valores. A única restrição é que exista uma constante A finita tal que $p(\theta|\mathbf{x}) < Aq(\theta)$. O método de rejeição consiste em gerar um valor θ^* da distribuição auxiliar q e aceitar este valor como sendo da distribuição a posteriori com probabilidade $p(\theta^*|\mathbf{x})/Aq(\theta^*)$. Caso contrário, θ^* não é aceito como um valor gerado da posteriori e o processo é repetido até que um valor seja aceito. O método também funciona se ao invés da posteriori, que em geral é desconhecida, usarmos a sua versão não normalizada, i.e $p(\mathbf{x}|\theta)p(\theta)$.

Podemos então usar o seguinte algoritmo,

1. gerar um valor θ^* da distribuição auxiliar;
2. gerar $u \sim U(0, 1)$;
3. se $u < p(\theta^*|\mathbf{x})/Aq(\theta^*)$ faça $\theta^{(j)} = \theta^*$, faça $j = j + 1$ e retorne ao passo 1. caso contrário retorne ao passo 1.

Tomando a priori $p(\theta)$ como densidade auxiliar a constante A deve ser tal que $p(\mathbf{x}|\theta) < A$. Esta desigualdade é satisfeita se tomarmos A como sendo o valor máximo da função de verossimilhança, i.e. $A = p(\mathbf{x}|\hat{\theta})$ onde $\hat{\theta}$ é o estimador de máxima verossimilhança de θ . Neste caso, a probabilidade de aceitação se simplifica para $p(\mathbf{x}|\theta)/p(\mathbf{x}|\hat{\theta})$.

Podemos então usar o seguinte algoritmo para gerar valores da posteriori,

1. gerar um valor θ^* da distribuição a priori;
2. gerar $u \sim U(0, 1)$;
3. aceitar θ^* como um valor da posteriori se $u < p(\mathbf{x}|\theta^*)/p(\mathbf{x}|\hat{\theta})$, caso contrário rejeitar θ^* e retornar ao passo 1.

Exemplo 4.4: Suponha que $X_1, \dots, X_n \sim N(\theta, 1)$ e assume-se uma distribuição a priori Cauchy(0,1) para θ . A função de verossimilhança é,

$$\begin{aligned} p(\mathbf{x}|\theta) &= \prod_{i=1}^n (2\pi)^{-1/2} \exp\left\{-\frac{(x_i - \theta)^2}{2}\right\} \propto \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} \end{aligned}$$

e o estimador de máxima verossimilhança é $\hat{\theta} = \bar{x}$. Usando o algoritmo acima, gera-se um valor da distribuição Cauchy(0,1) e a probabilidade de aceitação neste caso fica simplesmente $\exp[-n(\bar{x} - \theta)^2/2]$. A função do R a seguir obtém uma amostra de tamanho m de θ e como ilustração vamos gerar 50 observações da distribuição $N(2,1)$. Note que a taxa de aceitação foi extremamente baixa. Isto ocorreu devido ao conflito entre verossimilhança e priori.

```

> rej <- function(x, m) {
+   total = 0
+   theta = rep(0, m)
+   x.bar = mean(x)
+   n = length(x)
+   for (i in 1:m) {
+     accept = FALSE
+     while (!accept) {
+       total = total + 1
+       theta.new = rcauchy(1, 0, 1)
+       prob = exp(-0.5 * n * (theta.new - x.bar)^2)
+       u = runif(1, 0, 1)
+       if (u < prob) {
+         theta = c(theta, theta.new)
+         accept = TRUE
+       }
+     }
+   }
+   cat("\nTaxa de aceitacao", round(m/total, 4), "\n")
+   return(list(theta = theta, total = total))
+ }
```

> x = rnorm(n = 50, mean = 2, sd = 1)
> m = rej(x, m = 1000)

Taxa de aceitacao 0.0215

O problema é ilustrado na Figura 4.5 (gerada com os comandos abaixo) onde se pode notar que a maioria dos valores de θ foi gerada em regiões de baixa verossimilhança.

```

> x.bar = mean(x)
> x.sd = sd(x)
> curve(dnorm(x, x.bar, x.sd), xlab = expression(theta), from = -4,
+       to = 6, ylab = "", col = 1, lty = 1)
> curve(dcauchy(x, 0, 1), from = -4, to = 6, add = T, lty = 2)
> legend(-3, 0.4, legend = c("priori", "veross."), lty = c(2, 1))
> rug(m$theta)
```

Mudando a priori para Cauchy(2,1) obtém-se uma taxa de aceitação em torno de 10% o que ainda constitui uma amostra pequena. Na verdade o número de simulações deveria ser no mínimo 10000 neste caso.

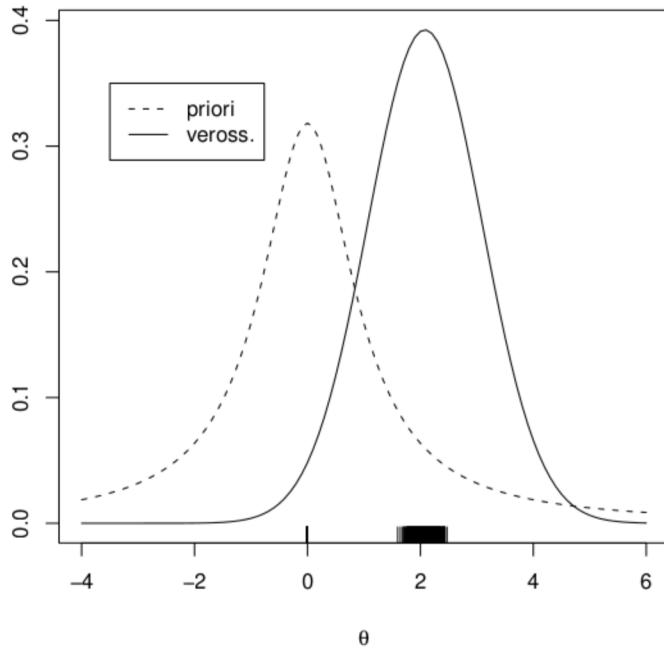


Figura 4.5: Verossimilhança normalizada e densidade a priori juntamente com valores simulados.

Portanto, um problema técnico associado ao método é a necessidade de se maximizar a função de verossimilhança o que pode não ser uma tarefa simples em modelos mais complexos. Se este for o caso então o método de rejeição perde o seu principal atrativo que é a simplicidade. Neste caso, o método da próxima seção passa a ser recomendado. Outro problema é que a taxa de aceitação pode ser muito baixa. Teremos que gerar muitos valores da distribuição auxiliar até conseguir um número suficiente de valores da distribuição a posteriori. Isto ocorrerá se as informações da distribuição a priori e da verossimilhança forem conflitantes já que neste caso os valores gerados terão baixa probabilidade de serem aceitos.

4.5.2 Reamostragem Ponderada

Estes métodos usam a mesma idéia de gerar valores de uma distribuição auxiliar porém sem a necessidade de maximização da verossimilhança. A desvantagem é que os valores obtidos são apenas aproximadamente distribuidos segundo a posteriori.

Suponha que temos uma amostra $\theta_1, \dots, \theta_n$ gerada da distribuição auxiliar q

e a partir dela construimos os pesos

$$w_i = \frac{p(\theta_i | \mathbf{x})/q(\theta_i)}{\sum_{j=1}^n p(\theta_j | \mathbf{x})/q(\theta_j)}, \quad i = 1, \dots, n$$

O método consiste em tomar uma segunda amostra (ou reamostra) de tamanho m da distribuição discreta em $\theta_1, \dots, \theta_n$ com probabilidades w_1, \dots, w_n . Aqui também não é necessário que se conheça completamente a posteriori mas apenas o produto priori vezes verossimilhança já que neste caso os pesos não se alteram.

Tomando novamente a priori como densidade auxiliar, i.e. $q(\theta) = p(\theta)$ os pesos se simplificam para

$$w_i = \frac{p(\mathbf{x} | \theta_i)}{\sum_{j=1}^n p(\mathbf{x} | \theta_j)}, \quad i = 1, \dots, n$$

e o algoritmo para geração de valores (aproximadamente) da posteriori então fica

1. gerar valores $\theta_1, \dots, \theta_n$ da distribuição a priori;
2. calcular os pesos $w_i, i = 1, \dots, n$;
3. reamostrar valores com probabilidades w_1, \dots, w_n .

Este método é essencialmente um *bootstrap* ponderado. O mesmo problema de informações conflitantes da priori e da verossimilhança pode ocorrer aqui. Neste caso, apenas poucos valores gerados da priori terão alta probabilidade de aparecerem na reamostra.

Exemplo 4.5: No Exemplo 4.4, utilizando reamostragem ponderada obtém-se os gráficos da Figura 4.6.

```
> reamostra <- function(x, n, m) {
+   x.bar = mean(x)
+   nobs = length(x)
+   theta = rcauchy(n, 0, 1)
+   peso = exp(-0.5 * nobs * (theta - x.bar)^2)
+   aux = sum(peso)
+   peso = peso/aux
+   theta.star = sample(theta, size = m, replace = TRUE, prob = peso)
+   return(list(amostra = theta, pesos = peso, reamostra = theta.star))
+ }
```

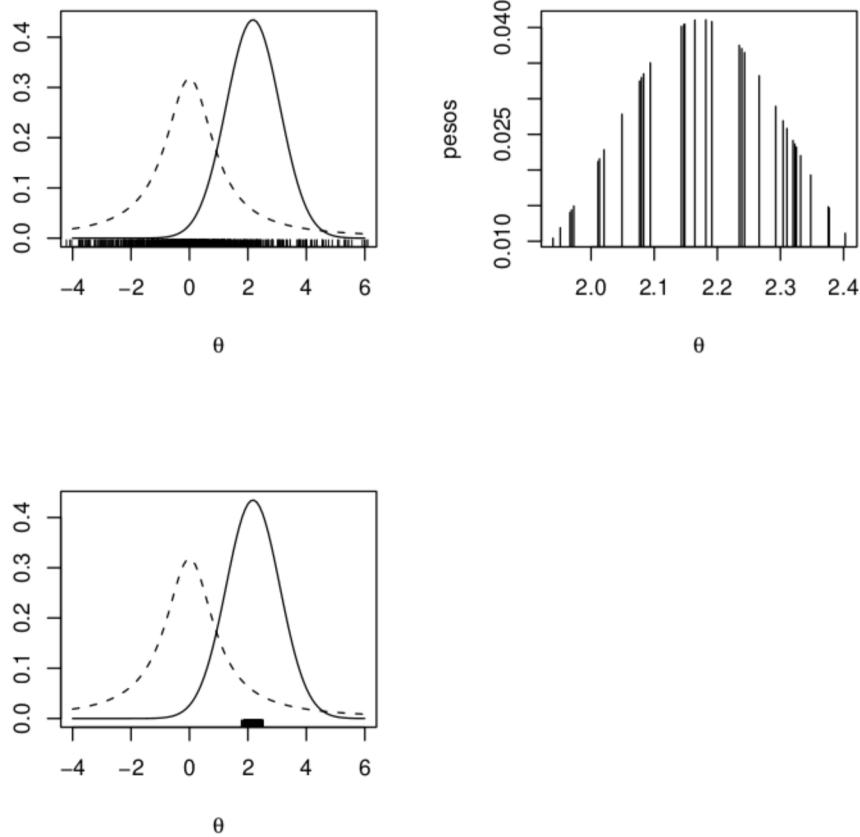


Figura 4.6: Verossimilhança normalizada (linha cheia), densidade a priori (linha tracejada) e os valores amostrados (a) e reamostrados (c). Em (b) os valores de θ com pesos maiores do que 0,01.

Exercícios

- Em um modelo de regressão linear simples temos que $y_i \sim N(\beta x_i, 1)$. Os dados observados são $\mathbf{y} = (-2, 0, 0, 0, 2)$ e $\mathbf{x} = (-2, -1, 0, 1, 2)$, e usamos uma priori vaga $N(0, 4)$ para β . Faça inferência sobre β obtendo uma amostra da posteriori usando reamostragem ponderada. Compare com a estimativa de máxima verossimilhança $\hat{\beta} = 0,8$.
- Para o mesmo modelo do exercício 1 e os mesmos dados suponha agora que a variância é desconhecida, i.e. $y_i \sim N(\beta x_i, \sigma^2)$. Usamos uma priori hierárquica para (β, σ^2) , i.e. $\beta | \sigma^2 \sim N(0, \sigma^2)$ e $\sigma^{-2} \sim G(0, 01, 0, 01)$.
 - Obtenha uma amostra da posteriori de (β, σ^2) usando reamostragem ponderada.

- (b) Baseado nesta amostra, faça um histograma das distribuições marginais de β e σ^2 .
- (c) Estime β e σ^2 usando uma aproximação para a média a posteriori. Compare com as estimativas de máxima verossimilhança.

4.6 Monte Carlo via cadeias de Markov

Em todos os métodos de simulação vistos até agora obtém-se uma amostra da distribuição a posteriori em um único passo. Os valores são gerados de forma independente e não há preocupação com a convergência do algoritmo, bastando que o tamanho da amostra seja suficientemente grande. Por isso estes métodos são chamados *não iterativos* (não confundir iteração com interação). No entanto, em muitos problemas pode ser bastante difícil, ou mesmo impossível, encontrar uma densidade de importância que seja simultaneamente uma boa aproximação da posteriori e fácil de ser amostrada.

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são uma alternativa aos métodos não iterativos em problemas complexos. A idéia ainda é obter uma amostra da distribuição a posteriori e calcular estimativas amostrais de características desta distribuição. A diferença é que aqui usaremos técnicas de simulação iterativa, baseadas em cadeias de Markov, e assim os valores gerados não serão mais independentes.

Nesta seção serão apresentados os métodos MCMC mais utilizados, o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. A idéia básica é simular um passeio aleatório no espaço de θ que converge para uma distribuição estacionária, que é a distribuição de interesse no problema. Uma discussão mais geral sobre o tema pode ser encontrada por exemplo em Gamerman (1997) e Gamerman & Lopes (2006).

4.6.1 Cadeias de Markov

Uma cadeia de Markov é um processo estocástico $\{X_0, X_1, \dots\}$ tal que a distribuição de X_t dados todos os valores anteriores X_0, \dots, X_{t-1} depende apenas de X_{t-1} . Matematicamente,

$$P(X_t \in A | X_0, \dots, X_{t-1}) = P(X_t \in A | X_{t-1})$$

para qualquer subconjunto A . Os métodos MCMC requerem ainda que a cadeia seja,

- homogênea, i.e. as probabilidades de transição de um estado para outro são invariantes;

- irreduzível, i.e. cada estado pode ser atingido a partir de qualquer outro em um número finito de iterações;
- aperiódica, i.e. não haja estados absorventes.

e os algoritmos que serão vistos aqui satisfazem a estas condições.

Suponha que uma distribuição $\pi(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$ seja conhecida a menos de uma constante multiplicativa porém complexa o bastante para não ser possível obter uma amostra diretamente. Dadas as realizações $\{\mathbf{X}^{(t)}, t = 0, 1, \dots\}$ de uma cadeia de Markov que tenha π como distribuição de equilíbrio então, sob as condições acima,

$$\mathbf{X}^{(t)} \xrightarrow{t \rightarrow \infty} \pi(x) \quad \text{e} \quad \frac{1}{n} \sum_{t=1}^n g(X_i^{(t)}) \xrightarrow{n \rightarrow \infty} \mathbb{E}_\pi(g(X_i)) \quad q.c.$$

Ou seja, embora a cadeia seja por definição dependente a média aritmética dos valores da cadeia é um estimador consistente da média teórica.

Uma questão importante de ordem prática é como os valores iniciais influenciam o comportamento da cadeia. A idéia é que conforme o número de iterações aumenta, a cadeia gradualmente *esquece* os valores iniciais e eventualmente converge para uma distribuição de equilíbrio. Assim, em aplicações práticas é comum que as iterações iniciais sejam descartadas, como se formassem uma *amostra de aquecimento*.

4.6.2 Acurácia Numérica

Na prática teremos um número finito de iterações e tomado

$$\hat{g} = \frac{1}{n} \sum_{t=1}^n g(X_i^{(t)})$$

como estimativa da $E(g(X_i))$ devemos calcular o seu erro padrão. Como a sequência de valores gerados é dependente pode-se mostrar que

$$Var(\hat{g}) = \frac{s^2}{n} \left[1 + 2 \sum_{k=1}^n \left(1 - \frac{k}{n} \right) \rho_k \right]$$

sendo s^2 a variância amostral e ρ_k a autocorrelação amostral de ordem k . Se $\rho_k > 0 \forall k$ então $Var(\hat{g}) > s^2/n$. Uma forma muito utilizada para o cálculo da variância do estimador é o método dos lotes aonde os valores da cadeia são divididos em k lotes de tamanho m e cada lote tem média B_i . O erro padrão de

\hat{g} é então estimado como

$$\sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (B_i - \bar{B})^2}$$

sendo m escolhido de modo que a correlação serial de ordem 1 entre as médias dos lotes seja menor do que 0,05.

Nas próximas seções serão apresentados e discutidos os algoritmos MCMC mais comumente utilizados.

4.6.3 Algoritmo de Metropolis-Hastings

Os algoritmos de Metropolis-Hastings usam a mesma idéia dos métodos de rejeição vistos no capítulo anterior, i.e. um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade. Este mecanismo de correção garante a convergência da cadeia para a distribuição de equilíbrio, que neste caso é a distribuição a posteriori.

Suponha que a cadeia esteja no estado θ e um valor θ' é gerado de uma *distribuição proposta* $q(\cdot|\theta)$. Note que a distribuição proposta pode depender do estado atual da cadeia, por exemplo $q(\cdot|\theta)$ poderia ser uma distribuição normal centrada em θ . O novo valor θ' é aceito com probabilidade

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta') q(\theta|\theta')}{\pi(\theta) q(\theta'|\theta)}\right). \quad (4.2)$$

onde π é a distribuição de interesse.

Uma característica importante é que só precisamos conhecer π parcialmente, i.e. a menos de uma constante já que neste caso a probabilidade (4.2) não se altera. Isto é fundamental em aplicações Bayesianas aonde não conhecemos completamente a posteriori. Note também que a cadeia pode permanecer no mesmo estado por muitas iterações e na prática costuma-se monitorar isto calculando a porcentagem média de iterações para as quais novos valores são aceitos.

Em termos práticos, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos,

1. Inicialize o contador de iterações $t = 0$ e especifique um valor inicial $\theta^{(0)}$.
2. Gere um novo valor θ' da distribuição $q(\cdot|\theta)$.
3. Calcule a probabilidade de aceitação $\alpha(\theta, \theta')$ e gere $u \sim U(0, 1)$.
4. Se $u \leq \alpha$ então aceite o novo valor e faça $\theta^{(t+1)} = \theta'$, caso contrário rejeite e faça $\theta^{(t+1)} = \theta$.

5. Incremente o contador de t para $t + 1$ e volte ao passo 2.

Embora a distribuição proposta possa ser escolhida arbitrariamente na prática deve-se tomar alguns cuidados para garantir a eficiência do algoritmo. Em aplicações Bayesianas a distribuição de interesse é a própria posteriori, i.e. $\pi = p(\theta|x)$ e a probabilidade de aceitação assume uma forma particular,

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{p(x|\theta') p(\theta') q(\theta|\theta')}{p(x|\theta) p(\theta) q(\theta'|\theta)} \right\}. \quad (4.3)$$

O algoritmo será ilustrado nos exemplos a seguir.

Exemplo 4.6 : Em uma certa população de animais sabe-se que cada animal pode pertencer a uma dentre 4 linhagens genéticas com probabilidades

$$p_1 = \frac{1}{2} + \frac{\theta}{4}, \quad p_2 = \frac{1-\theta}{4}, \quad p_3 = \frac{1-\theta}{4}, \quad p_4 = \frac{\theta}{4}.$$

sendo $0 < \theta < 1$ um parâmetro desconhecido. Para qualquer $\theta \in (0, 1)$ é fácil verificar que $p_i > 0$, $i = 1, 2, 3, 4$ e $p_1 + p_2 + p_3 + p_4 = 1$. Observando-se n animais dentre os quais y_i pertencem à linhagem i então o vetor aleatório $\mathbf{Y} = (y_1, y_2, y_3, y_4)$ tem distribuição multinomial com parâmetros n, p_1, p_2, p_3, p_4 e portanto,

$$\begin{aligned} p(\mathbf{y}|\theta) &= \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4} \\ &\propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}. \end{aligned}$$

Atribuindo a distribuição a priori $\theta \sim U(0, 1)$ segue que a densidade a posteriori é proporcional à expressão acima. Então,

$$p(\theta|\mathbf{y}) \propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}.$$

Tomando a distribuição $U(0, 1)$ como proposta então $q(\theta) = 1$, $\forall \theta$ e a probabilidade (4.3) se simplifica para

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{p(x|\theta')}{p(x|\theta)} \right\} = \min \left\{ 1, \left(\frac{2+\theta'}{2+\theta} \right)^{y_1} \left(\frac{1-\theta'}{1-\theta} \right)^{y_2+y_3} \left(\frac{\theta'}{\theta} \right)^{y_4} \right\}.$$

Podemos programar este algoritmo com os comandos do R a seguir.

```
> p <- function(x, y) {
+   (2 + x)^y[1] * (1 - x)^(y[2] + y[3]) * x^y[4]
+ }
```

```

> metr0 <- function(n, y, fun, start) {
+   theta = c(start, rep(NA, n - 1))
+   taxa = 0
+   for (i in 2:n) {
+     x = runif(1)
+     A = fun(x, y)/fun(theta[i - 1], y)
+     prob = min(1, A)
+     if (runif(1) < prob) {
+       theta[i] = x
+       taxa = taxa + 1
+     }
+     else {
+       theta[i] = theta[i - 1]
+     }
+   }
+   return(list(theta = theta, taxa = taxa/n))
+ }
```

Suponha que foram observados 197 animais com os números de animais nas categorias dados por $\mathbf{y} = (125, 18, 20, 34)$ e foi gerada uma cadeia de Markov com 2000 valores de θ . Os valores simulados e as primeiras 30 autocorrelações amostrais de θ estão na Figura 4.7. A cadeia parece ter convergido após algumas iterações e podemos descartar os 100 primeiros valores (esta foi a nossa amostra de aquecimento). Note também que a cadeia é altamente correlacionada ao longo das iterações e isto é devido a alta taxa de rejeição por causa da escolha de q .

```

> y = c(125, 18, 20, 34)
> n = 2000
> m = metr0(n, y, fun = p, start = 0.5)
> m$taxa

[1] 0.17
```

Dada uma amostra com valores de θ temos também amostras de valores de (p_1, p_2, p_3, p_4) que podem ser resumidas da seguinte forma,

```

> p1 = m$theta/4 + 0.5
> p2 = (1 - m$theta)/4
> p3 = p2
> p4 = m$theta/4
> z = as.mcmc(cbind(p1, p2, p3, p4))
> colnames(z) = c("p1", "p2", "p3", "p4")
> b = summary(window(z, start = 501))
> print(b, digits = 3)
```

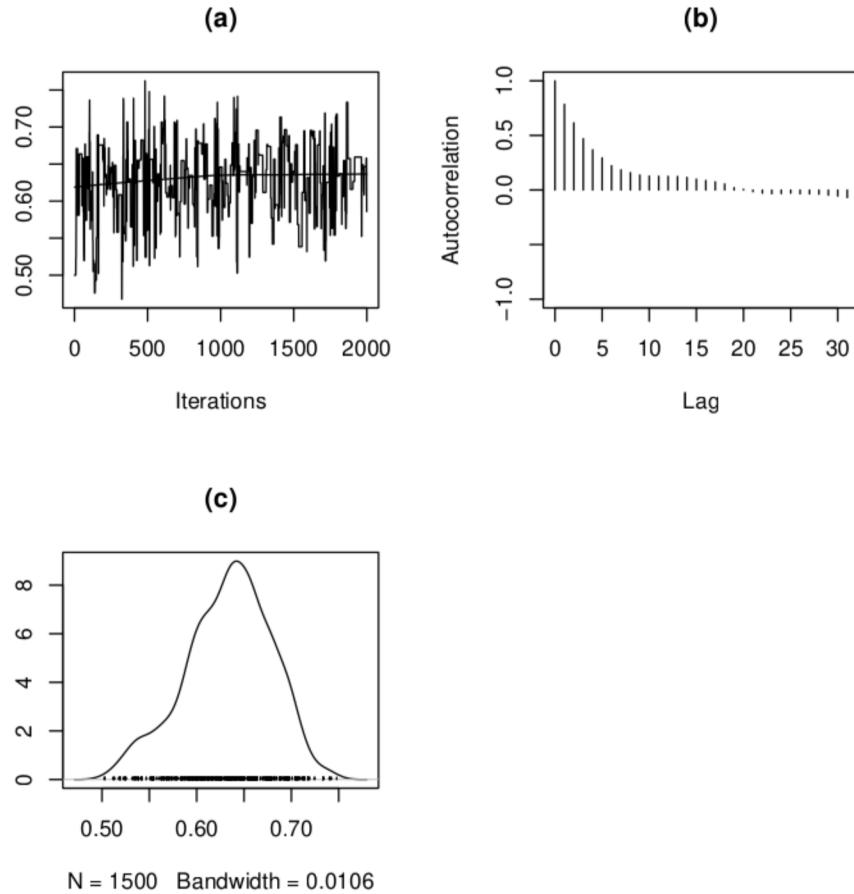


Figura 4.7: (a) 2000 valores simulados de θ , (b) 30 primeiras autocorrelações amostrais após aquecimento, (c) Densidade a posteriori estimada.

```

Iterations = 501:2000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1500

```

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
p1	0.6584	0.0114	0.000294	0.000954
p2	0.0916	0.0114	0.000294	0.000954
p3	0.0916	0.0114	0.000294	0.000954
p4	0.1584	0.0114	0.000294	0.000954

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
p1	0.6340	0.6512	0.6592	0.6656	0.678
p2	0.0721	0.0844	0.0908	0.0988	0.116
p3	0.0721	0.0844	0.0908	0.0988	0.116
p4	0.1340	0.1512	0.1592	0.1656	0.178

Exemplo 4.7: Suponha que queremos simular valores $X \sim N(0, 1)$ propondo valores $Y \sim N(x, \sigma^2)$. Neste caso as densidades propostas no numerador e denominador de (4.2) se cancelam e a probabilidade de aceitação fica

$$\alpha(x, y) = \min \left\{ 1, \exp \left(-\frac{1}{2}(y^2 - x^2) \right) \right\}.$$

Fixando os valores $\sigma = 0.5$ e $\sigma = 10$ foram simuladas as cadeias que aparecem na Figura 4.8. Note que o valor de σ teve um grande impacto na taxa de aceitação do algoritmo. Isto ocorre porque com $\sigma = 0.5$ a distribuição proposta está muito mais próxima da distribuição de interesse do que com $\sigma = 10$.

```

> metrop <- function(n, sigma) {
+   x = c(0, rep(NA, n - 1))
+   for (i in 2:n) {
+     y = rnorm(1, x[i - 1], sigma)
+     prob = min(1, exp(-0.5 * (y^2 - x[i - 1]^2)))
+     u = runif(1, 0, 1)
+     if (u < prob)
+       x[i] = y
+     else x[i] = x[i - 1]
+   }
+   return(x)
+ }

```

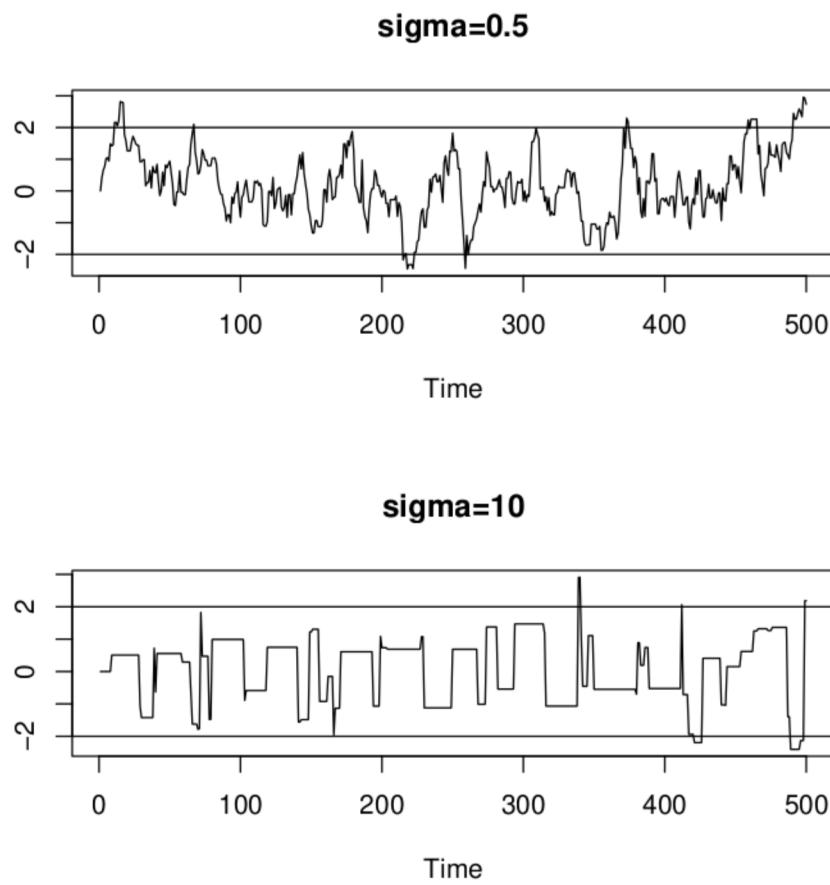


Figura 4.8: 500 valores simulados para o Exemplo 4.7 usando o algoritmo de Metropolis-Hastings com (a) $\sigma = 0.5$ e (b) $\sigma = 10$.

Nos Exemplos 4.6 e 4.7 foram ilustrados casos especiais do algoritmo nos quais a distribuição proposta não depende do estado atual ou a dependência é na forma de um passeio aleatório. Estes casos são formalizados a seguir.

4.6.4 Casos Especiais

Um caso particular é quando a distribuição proposta não depende do estado atual da cadeia, i.e. $q(\theta'|\theta) = q(\theta')$. Em geral, $q(\cdot)$ deve ser uma boa aproximação de $\pi(\cdot)$, mas é mais seguro se $q(\cdot)$ tiver caudas mais pesadas do que $\pi(\cdot)$. A probabilidade de aceitação agora fica,

$$\alpha(\theta, \theta') = \min \left\{ 1, \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\theta)}{q(\theta')} \right\}. \quad (4.4)$$

Note que embora os valores θ' sejam gerados de forma independente a cadeia resultante não será i.i.d. já que a probabilidade de aceitação ainda depende de θ .

Outro caso particular é chamado algoritmo de Metropolis e considera apenas propostas simétricas, i.e., $q(\theta'|\theta) = q(\theta|\theta')$ para todos os valores de θ e θ' . Neste caso a probabilidade de aceitação se reduz para

$$\alpha(\theta, \theta') = \min \left(1, \frac{\pi(\theta')}{\pi(\theta)} \right).$$

Um algoritmo de Metropolis muito utilizado é baseado em um passeio aleatório de modo que a probabilidade da cadeia mover-se de θ para θ' depende apenas da distância entre eles, i.e. $q(\theta'|\theta) = q(|\theta - \theta'|)$. Neste caso, se usarmos uma distribuição proposta com variância σ^2 duas situações extremas podem ocorrer,

1. se σ^2 for muito pequena os valores gerados estarão próximos do valor atual e quase sempre serão aceitos. Mas levará muitas iterações até o algoritmo cobrir todo o espaço do parâmetro;
2. valores grandes de σ^2 levam a uma taxa de rejeição excessivamente alta e a cadeia se movimenta muito pouco.

Nas duas situações o algoritmo fica ineficiente e na prática temos que tentar vários valores de σ^2 .

De um modo geral $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$ será um vetor de parâmetros de dimensão d . Neste caso, pode ser computacionalmente mais eficiente dividir $\boldsymbol{\theta}$ em k blocos $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k\}$ e dentro de cada iteração teremos o algoritmo aplicado k vezes. Definindo o vetor $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \dots, \boldsymbol{\theta}_k)$ que contém todos os elementos de $\boldsymbol{\theta}$ exceto $\boldsymbol{\theta}_i$ suponha que na iteração $t + 1$ os blocos $1, 2, \dots, i - 1$ já foram atualizados, i.e.

$$\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1^{(t+1)}, \dots, \boldsymbol{\theta}_{i-1}^{(t+1)}, \boldsymbol{\theta}_{i+1}^{(t)}, \dots, \boldsymbol{\theta}_k^{(t)}).$$

Para atualizar a i -ésima componente, um valor de θ_i é gerado da distribuição proposta $q(\cdot|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i})$ e este valor candidato é aceito com probabilidade

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}'_i | \boldsymbol{\theta}_{-i}) q(\boldsymbol{\theta}_i | \boldsymbol{\theta}'_i, \boldsymbol{\theta}_{-i})}{\pi(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}) q(\boldsymbol{\theta}'_i | \boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i})} \right\}. \quad (4.5)$$

Aqui, $\pi(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i})$ é chamada de *distribuição condicional completa* como será visto na próxima seção.

Exercícios

1. Assumindo que a distribuição estacionária é $N(0, 1)$,
 - (a) faça 500 iterações do algoritmo de Metropolis com distribuições propostas $N(\theta; 0, 5)$, $N(\theta; 0, 1)$ e $N(\theta, 10)$.
 - (b) faça os gráficos dos valores das cadeias ao longo das iterações. Existe alguma indicação de convergência nos gráficos?
 - (c) Calcule as taxas de aceitação.
2. Suponha que a distribuição estacionária é $N(0, 1)$.
 - (a) Para distribuições propostas Cauchy($0, \sigma$), selecione experimentalmente o valor de σ que maximiza a taxa de aceitação.
 - (b) Para este valor de σ faça os gráficos dos valores simulados da cadeia ao longo das iterações e verifique se há indicação de convergência.
 - (c) Repita os itens anteriores com a distribuição proposta Cauchy(θ, σ).

4.6.5 Amostrador de Gibbs

No amostrador de Gibbs a cadeia irá sempre se mover para um novo valor, i.e não existe mecanismo de aceitação-rejeição. As transições de um estado para outro são feitas de acordo com as *distribuições condicionais completas* $\pi(\theta_i | \boldsymbol{\theta}_{-i})$, onde $\boldsymbol{\theta}_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d)'$.

Em geral, cada uma das componentes θ_i pode ser uni ou multidimensional. Portanto, a distribuição condicional completa é a distribuição da i -ésima componente de $\boldsymbol{\theta}$ condicionada em todas as outras componentes. Ela é obtida a partir da distribuição conjunta como,

$$\pi(\theta_i | \boldsymbol{\theta}_{-i}) = \frac{\pi(\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}) d\theta_i}.$$

Assim, para obter a distribuição condicional completa de x_i basta pegar os termos da distribuição conjunta que não dependem de x_i .

Exemplo 4.8 : Em um modelo Bayesiano para os dados \mathbf{y} que depende dos parâmetros θ , λ e δ suponha que a distribuição conjunta é dada por

$$p(\mathbf{y}, \theta, \lambda, \delta) \propto p(\mathbf{y}|\theta, \delta)p(\theta|\lambda)p(\lambda)p(\delta).$$

Após observar \mathbf{y} as distribuições a posteriori de cada parâmetro dados todos os outros são

$$\begin{aligned}\pi(\theta|\mathbf{y}, \lambda, \delta) &\propto p(\mathbf{y}|\theta, \delta)p(\theta|\lambda) \\ \pi(\lambda|\mathbf{y}, \theta, \delta) &\propto p(\theta|\lambda)p(\lambda) \\ \pi(\delta|\mathbf{y}, \theta, \lambda) &\propto p(\mathbf{y}|\theta, \delta)p(\delta).\end{aligned}$$

Em muitas situações, a geração de uma amostra diretamente de $\pi(\boldsymbol{\theta})$ pode ser custosa, complicada ou simplesmente impossível. Mas se as distribuições condicionais completas forem completamente conhecidas, então o amostrador de Gibbs é definido pelo seguinte esquema,

1. inicialize o contador de iterações da cadeia $t = 0$;
2. especifique valores iniciais $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})'$;
3. obtenha um novo valor de $\boldsymbol{\theta}^{(t)}$ a partir de $\boldsymbol{\theta}^{(t-1)}$ através da geração sucessiva dos valores

$$\begin{aligned}\theta_1^{(t)} &\sim \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)}) \\ \theta_2^{(t)} &\sim \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)}) \\ &\vdots \\ \theta_d^{(t)} &\sim \pi(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{d-1}^{(t)})\end{aligned}$$

4. Incremente o contador de t para $t + 1$ e retorne ao passo 2 até obter convergência.

Assim, cada iteração se completa após d movimentos ao longo dos eixos coordenados das componentes de $\boldsymbol{\theta}$. Após a convergência, os valores resultantes formam uma amostra de $\pi(\boldsymbol{\theta})$. Vale notar que, mesmo em problema de grandes dimensões todas as simulações podem ser univariadas, o que em geral é uma vantagem computacional.

Note também que o amostrador de Gibbs é um caso especial do algoritmo de Metropolis-Hastings, no qual os elementos de $\boldsymbol{\theta}$ são atualizados um de cada vez

(ou em blocos), tomando a distribuição condicional completa como proposta e probabilidade de aceitação igual a 1.

Mais detalhes sobre o amostrado de Gibbs e outros algoritmos relacionados podem ser obtidos, por exemplo, em Gamerman (1997, Cap. 5) e Robert & Casella (1999, Cap. 7).

Exemplo 4.9: Suponha que $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$ com μ e σ^2 desconhecidos. Definindo $\tau = \sigma^{-2}$ a função de verossimilhança é dada por

$$p(\mathbf{y}|\mu, \tau) \propto \tau^{n/2} \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right]$$

e especificando prioris independentes $\mu \sim N(0, s^2)$, sendo s^2 a variância amostral e $\tau \sim Gama(a, b)$, com a e b conhecidos, segue que

$$\begin{aligned} p(\mu, \tau | \mathbf{y}) &\propto p(\mathbf{y}|\mu, \tau)p(\mu)p(\tau) \\ &\propto \tau^{n/2} \exp \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2 \right] \exp \left[-\frac{\mu^2}{2s^2} \right] \tau^{a-1} e^{-b\tau}. \end{aligned}$$

Esta distribuição conjunta não tem forma padrão mas as condicionais completas são fáceis de obter,

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

onde $C^{-1} = n\tau + s^{-2}$ e $m = C\bar{y}$ e

$$p(\tau | \mathbf{y}, \mu) \propto \tau^{a+n/2-1} \exp \left[-\tau \left(b + \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2 \right) \right].$$

Segue então que

$$\begin{aligned} \mu | \mathbf{y}, \tau &\sim N(m, C) \\ \tau | \mathbf{y}, \mu &\sim Gama \left(a + \frac{n}{2}, b + \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2 \right) \end{aligned}$$

e o amostrador de Gibbs pode ser implementado facilmente gerando valores destas distribuições alternadamente.

Exemplo 4.10: Em um processo de contagem no qual foram observados

Y_1, \dots, Y_n suspeita-se que houve um ponto de mudança m tal que

$$\begin{aligned} Y_i &\sim Poisson(\lambda), \quad i = 1, \dots, m \\ Y_i &\sim Poisson(\phi), \quad i = m + 1, \dots, n. \end{aligned}$$

O objetivo é estimar o ponto de mudança m e os parâmetros dos 2 processos de Poisson. Assumindo-se as distribuições a priori independentes

$$\begin{aligned} \lambda &\sim Gama(a, b) \\ \phi &\sim Gama(c, d) \\ m &\sim Uniforme\{1, \dots, n\} \end{aligned}$$

a densidade a posteriori fica

$$\begin{aligned} p(\lambda, \phi, m | \mathbf{y}) &\propto \prod_{i=1}^m e^{-\lambda} \lambda^{y_i} \prod_{i=m+1}^n e^{-\phi} \phi^{y_i} \lambda^{a-1} e^{-b\lambda} \phi^{c-1} e^{-d\phi} \frac{1}{n} \\ &\propto \lambda^{a+t_1-1} e^{-(b+m)\lambda} \phi^{c+t_2-1} e^{-(d+n-m)\phi} \frac{1}{n} \end{aligned}$$

sendo $t_1 = \sum_{i=1}^m y_i$ e $t_2 = \sum_{i=m+1}^n y_i$. Neste caso não é difícil verificar que as distribuições condicionais completas ficam

$$\begin{aligned} p(\lambda | \phi, m, \mathbf{y}) &\propto \lambda^{a+t_1-1} e^{-(b+m)\lambda} \quad \text{ou} \quad \lambda | \phi, m, \mathbf{y} \sim Gama(a + t_1, b + m) \\ p(\phi | \lambda, m, \mathbf{y}) &\propto \phi^{c+t_2-1} e^{-(d+n-m)\phi} \quad \text{ou} \quad \phi | \lambda, m, \mathbf{y} \sim Gama(c + t_2, d + n - m) \\ p(m | \lambda, \phi, \mathbf{y}) &\propto \lambda^{t_1} e^{-m\lambda} \phi^{t_2} e^{-(n-m)\phi}, \quad m = 1, \dots, n. \end{aligned}$$

A função do R abaixo obtém uma amostra da posteriori conjunta simulando valores destas condicionais completas.

```
> Gibbs <- function(a, b, c, d, y, niter) {
+   N = length(y)
+   lambda = phi = m = matrix(0, nrow = niter)
+   lambda[1] = 1
+   phi[1] = 1
+   m[1] = 10
+   for (i in 2:niter) {
+     t1 = sum(y[1:m[i - 1]])
+     t2 = 0
+     if (m[i - 1] < N)
+       t2 = sum(y[(m[i - 1] + 1):N])
+     lambda[i] = rgamma(1, (a + t1), (b + m[i - 1]))
+     phi[i] = rgamma(1, (c + t2), (d + N - m[i - 1]))
+     prob = NULL
+     for (j in 1:N) {
+       t1 = sum(y[1:j])
+       t2 = 0
+       if (j < N) {
+         t2 = sum(y[(j + 1):N])
+       }
+       aux = (lambda[i]^t1) * exp(-j * lambda[i]) * (phi[i]^t2) *
+             exp(-(N - j) * phi[i])
+       prob = c(prob, aux)
+     }
+     soma = sum(prob)
+     probm = prob/soma
+     m[i] = sample(x = N, size = 1, prob = probm)
+   }
+   return(list(lambda = lambda, phi = phi, m = m))
+ }
```

Testando a função Gibbs com 40 dados simulados de processos com médias 2 e 5 e ponto de mudança 23.

```
> y = c(rpois(n = 22, lambda = 2), rpois(n = 18, lambda = 5))
> x = Gibbs(a = 0.1, b = 0.1, c = 0.1, d = 0.1, y = y, niter = 2000)
```

Podemos usar o pacote `coda` para analisar os valores simulados. As 1000 primeiras simulações são descartadas como amostra de aquecimento.

```
> library(coda)
> amostra = cbind(x$lambda, x$phi, x$m)[1001:2000, ]
```

```
> theta = mcmc(amostra)
> colnames(theta) = names(x)
> summary(theta)
```

Iterations = 1:1000
 Thinning interval = 1
 Number of chains = 1
 Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
 plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
lambda	2.273	0.3247	0.01027	0.00865
phi	5.246	0.5569	0.01761	0.02049
m	21.612	1.6125	0.05099	0.06403

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
lambda	1.668	2.054	2.258	2.479	2.979
phi	4.213	4.843	5.230	5.610	6.398
m	18.975	21.000	22.000	22.000	24.025

A partir dos valores simulados de m podemos estimar suas probabilidades,

```
> tm = table(theta[, 3])/1000
> print(tm)
```

	7	8	9	10	11	14	15	16	17	18	19	20	21
0.001	0.001	0.001	0.001	0.001	0.005	0.002	0.004	0.001	0.007	0.012	0.059	0.196	
22	23	24	25	26	27								
0.648	0.010	0.025	0.010	0.013	0.002								

Finalmente, pode-se estimar as contagens médias condicionando nos valor de m com maior probabilidade.

```
> lambda.22 = theta[, 1][theta[, 3] == 22]
> phi.22 = theta[, 2][theta[, 3] == 22]
> theta.22 = as.mcmc(cbind(lambda.22, phi.22))
```

```
> plot(theta)
```

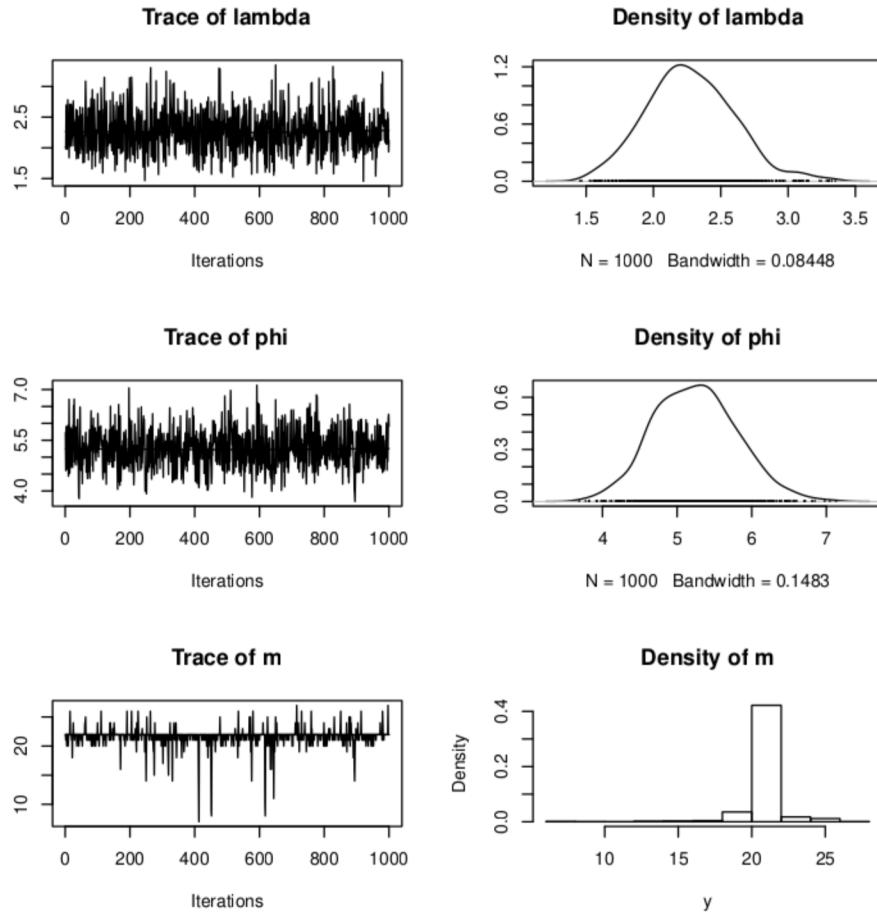


Figura 4.9: rtwert

4.7 Problemas de Dimensão Variável

Em muitas aplicações práticas é razoável assumir que existe incerteza também em relação ao modelo que melhor se ajusta a um conjunto de dados. Do ponto de vista Bayesiano esta incerteza é simplesmente incorporada ao problema de inferência considerando-se o próprio modelo como mais um parâmetro desconhecido a ser estimado. Assim os diferentes modelos terão uma distribuição de probabilidades.

Para isto vamos criar uma variável aleatória discreta k que funciona como indicador de modelo e atribuir probabilidades a priori $p(k)$ para cada modelo. Além disso, para cada k existe um vetor de parâmetros $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^{n_k}$ com

- uma verossimilhança $p(\mathbf{y}|\boldsymbol{\theta}^{(k)}, k)$
- uma distribuição a priori $p(\boldsymbol{\theta}^{(k)}|k)$.

```
> plot(tm)
```

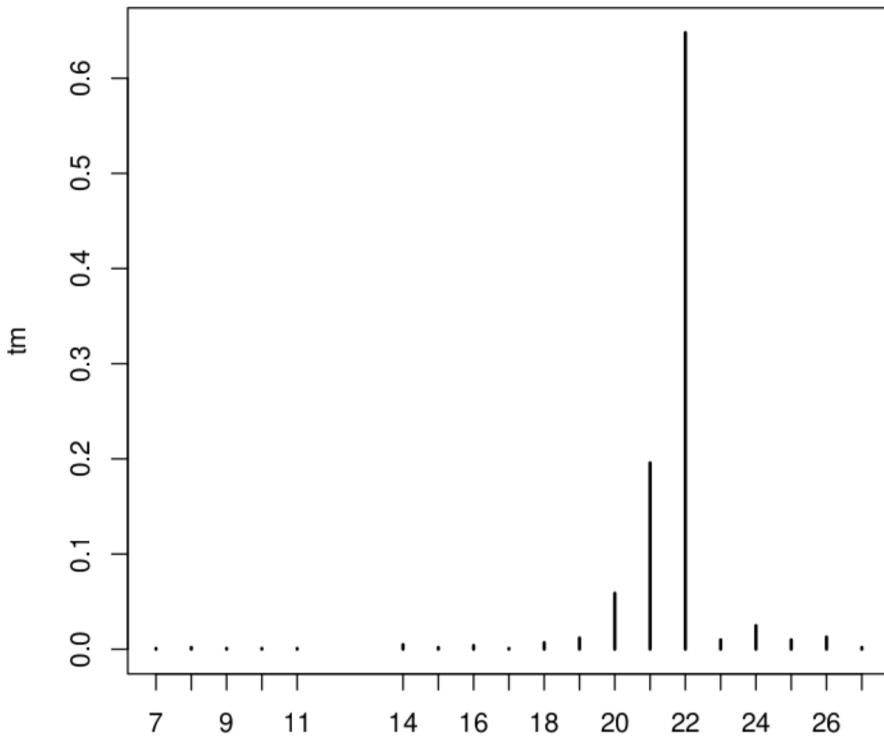


Figura 4.10:

Se M é conjunto de todos os possíveis modelos (ou modelos candidatos), então as probabilidades a posteriori de cada possível modelo são dadas por

$$\pi(k|\mathbf{y}) = \frac{p(k) p(\mathbf{y}|k)}{\sum_{k \in M} p(k) p(\mathbf{y}|k)}, \quad k \in M$$

sendo $p(\mathbf{y}|k)$ a *verossimilhança marginal* obtida como

$$p(\mathbf{y}|k) = \int p(\mathbf{y}|\boldsymbol{\theta}, k) p(\boldsymbol{\theta}|k) d\boldsymbol{\theta}.$$

O problema aqui é que esta última integral só é analiticamente tratável em alguns casos restritos. Além disso, se o número de modelos candidatos for muito grande calcular (ou aproximar) $p(\mathbf{y}|k)$ pode ser inviável na prática.

```
> plot(theta.22)
```

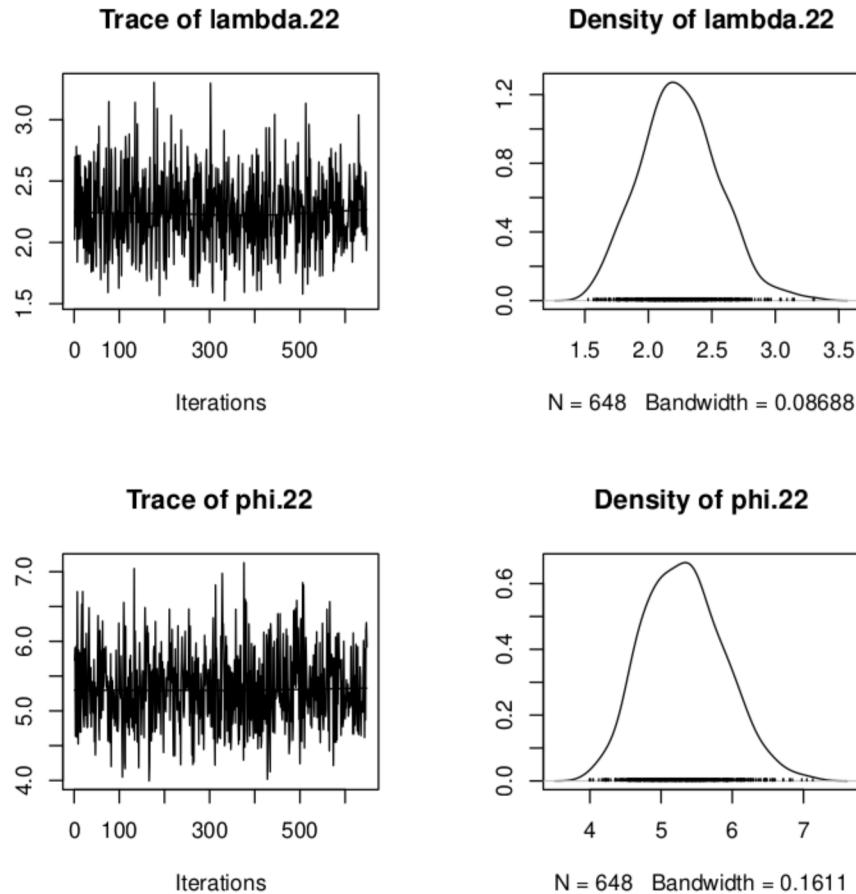


Figura 4.11:

Por outro lado, se for especificada a distribuição de interesse como a seguinte posteriori conjunta,

$$\pi(\boldsymbol{\theta}, k | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}, k) p(\boldsymbol{\theta} | k) p(k)$$

e conseguirmos simular valores desta distribuição então automaticamente teremos uma amostra aproximada de $\pi(k | \mathbf{y})$ e $\pi(\boldsymbol{\theta} | k, \mathbf{y})$.

Note que neste caso estamos admitindo que a dimensão de $\boldsymbol{\theta}$ pode variar ao longo dos modelos e precisamos então construir uma cadeia com espaço de estados que muda de dimensão ao longo das iterações. Os algoritmos de Metropolis-Hastings e o amostrador de Gibbs não podem ser utilizados já que são definidos apenas para distribuições com dimensão fixa. Embora existam outras possibilidades iremos estudar os algoritmos MCMC com saltos reversíveis (Green 1995) que são particularmente úteis no contexto de seleção Bayesiana de modelos.

4.7.1 MCMC com Saltos Reversíveis (RJMCMC)

Este algoritmo é baseado na abordagem usual dos métodos de Metropolis-Hastings de propor um novo valor para a cadeia e definir uma probabilidade de aceitação. No entanto, os movimentos podem ser entre espaços de dimensões diferentes como veremos a seguir. Em cada iteração o algoritmo envolve a atualização dos parâmetros, dado o modelo, usando os métodos MCMC usuais discutidos anteriormente e a atualização da dimensão usando o seguinte procedimento.

Suponha que o estado atual da cadeia é $(k, \boldsymbol{\theta})$, i.e. estamos no modelo k com parâmetros $\boldsymbol{\theta}$ e um novo modelo k' com parâmetros $\boldsymbol{\theta}'$ é proposto com probabilidade $r_{k,k'}$. Em geral isto significa incluir ou retirar parâmetros do modelo atual. Vamos assumir inicialmente que o modelo proposto tem dimensão maior, i.e. $n_{k'} > n_k$ e que $\boldsymbol{\theta}' = g(\boldsymbol{\theta}, \mathbf{u})$ para uma função determinística g e um vetor aleatório $\mathbf{u} \sim q(\mathbf{u})$ com dimensão $n_{k'} - n_k$. Então o seguinte algoritmo é utilizado,

- proponha $(k, \boldsymbol{\theta}) \rightarrow (k', \boldsymbol{\theta}')$ com probabilidade $r_{k,k'}$
- gere $\mathbf{u} \sim q(\mathbf{u})$ com dimensão $n_{k'} - n_k$
- faça $\boldsymbol{\theta}' = g(\boldsymbol{\theta}, \mathbf{u})$,
- aceite $(k', \boldsymbol{\theta}')$ com probabilidade $\min(1, A)$ sendo

$$A = \frac{\pi(k', \boldsymbol{\theta}')}{\pi(k, \boldsymbol{\theta})} \times \frac{r_{k',k}}{r_{k,k'} q(\mathbf{u})} \left| \frac{\partial g(\boldsymbol{\theta}, \mathbf{u})}{\partial (\boldsymbol{\theta}, \mathbf{u})} \right|.$$

Exemplo 4.11 : Sejam Y_1, \dots, Y_n os tempos de vida de componentes eletrônicos sorteados ao acaso e existe incerteza em relação a distribuição dos dados. Sabe-se que

$$Y_i \sim \text{Exp}(\lambda) \text{ (Modelo 1)} \quad \text{ou} \quad Y_i \sim \text{Gama}(\alpha, \beta) \text{ (Modelo 2)}, \quad i = 1, \dots, n.$$

Suponha que atribuimos as probabilidades a priori $p(k) = 1/2$ para o indicador de modelo e as seguintes distribuições a priori foram atribuídas aos parâmetros dentro de cada modelo,

$$\lambda|k=1 \sim \text{Gama}(2, 1) \quad \alpha|k=2 \sim \text{Gama}(4, 2) \quad \text{e} \quad \beta|k=2 \sim \text{Gama}(4, 2).$$

Dado o modelo, as funções de verossimilhança ficam

$$p(\mathbf{y}|\lambda, k=1) = \lambda^n e^{-\lambda \sum y_i}$$

$$p(\mathbf{y}|\alpha, \beta, k=2) = \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \prod y_i^{\alpha-1} e^{-\beta \sum y_i}$$

as distribuições condicionais completas são facilmente obtidas como

$$\begin{aligned}\lambda|\mathbf{y}, \alpha, \beta, k=1 &\sim Gama(n+2, 1 + \sum y_i) \\ \beta|\mathbf{y}, \alpha, \lambda, k=2 &\sim Gama(n\alpha+4, 2 + \sum y_i) \\ p(\alpha|\mathbf{y}, \beta, \lambda, k=2) &\propto \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \prod y_i^{\alpha-1} \alpha^3 e^{-2\alpha}\end{aligned}$$

A distribuição condicional completa de α não é conhecida então vamos usar o algoritmo de Metropolis-Hastings propondo valores $\alpha' \sim U[\alpha - \epsilon, \alpha + \epsilon]$. A função a seguir atualiza o valor de α segundo este esquema.

```
> mh.alpha <- function(y, n, alpha, beta, eps) {
+   z = runif(1, alpha - eps, alpha + eps)
+   if (z <= 0) {
+     acc = 0
+   } else {
+     t1 = prod(y)
+     num = beta^(n * z) * t1^(z - 1)/(gamma(z)^n)
+     den = beta^(n * alpha) * t1^(alpha - 1)/(gamma(alpha)^n)
+     num = num * exp(-2 * z) * z^3
+     den = den * exp(-2 * alpha) * alpha^3
+   }
+   aceita = min(1, num/den)
+   u = runif(1)
+   newalpha = ifelse(u < aceita, z, alpha)
+   return(newalpha)
+ }
```

Suponha que o modelo atual é $Exp(\lambda)$ e queremos propor o modelo $Gama(\alpha, \beta)$. Um possível esquema de atualização é o seguinte,

1. gere $u \sim Gama(a, b)$
2. defina $(\alpha, \beta) = g(\lambda, u) = (u, \lambda u)$
3. calcule o Jacobiano,

$$\begin{vmatrix} 0 & 1 \\ u & \lambda \end{vmatrix} = u$$

4. aceite o novo modelo com probabilidade $\min(1, A)$ sendo

$$A = \frac{p(\mathbf{y} | \alpha, \beta, k=2)}{p(\mathbf{y} | \lambda, k=1)} \frac{p(\alpha)p(\beta)}{p(\lambda)} \frac{u}{q(u)}$$

Note que transformação no item (2) preserva a média, ou seja $E(Y) = 1/\lambda$ sob o modelo exponencial e $E(Y) = u/\lambda u = 1/\lambda$ sob o modelo gama.

Se o modelo atual for $Gama(\alpha, \beta)$ e propomos o modelo $Exp(\lambda)$ o esquema reverso consiste em fazer $(\lambda, u) = g^{-1}(\alpha, \beta) = (\beta/\alpha, \alpha)$. A probabilidade de aceitação é simplesmente $\min(1, 1/A)$ substituindo $u = \alpha$.

```
> rj.modelo <- function(y, n, lambda, alpha, beta, model, a, b) {
+   if (model == 1) {
+     u = rgamma(1, a, b)
+     alpha1 = u
+     beta1 = lambda * u
+     lambda1 = lambda
+   }
+   else {
+     lambda1 = beta/alpha
+     alpha1 = alpha
+     beta1 = beta
+     u = alpha
+   }
+   t1 = prod(y)
+   t2 = sum(y)
+   num = beta1^(n * alpha1) * t1^(alpha1 - 1) * exp(-beta1 *
+     t2)/(gamma(alpha1)^n)
+   num = num * 2^4 * alpha1^3 * exp(-2 * alpha1)/gamma(4)
+   num = num * 2^4 * beta1^3 * exp(-2 * beta1)/gamma(4) * alpha1
+   den = (lambda1^n) * exp(-lambda1 * t2)
+   den = den * lambda1 * exp(-lambda1)/gamma(2)
+   den = den * b^a * u^(a - 1) * exp(-b * u)/gamma(a)
+   u = runif(1, 0, 1)
+   if (model == 1) {
+     aceita = min(1, num/den)
+     if (u < aceita) {
+       model = 2
+       alpha = alpha1
+       beta = beta1
+     }
+   }
}
```

```

+   else {
+     aceita = min(1, den/num)
+     if (u < aceita) {
+       model = 1
+       lambda = lambda1
+     }
+   }
+   if (model == 1)
+     return(list(model = model, lambda = lambda))
+   else return(list(model = model, alpha = alpha, beta = beta))
+ }
```

Finalmente o algoritmo pode ser implementado para atualizar tanto o modelo quanto os parâmetros dentro do modelo.

```

> rjcmc <- function(niter, nburn, y, n, a, b, eps = 0.25) {
+   x = matrix(0, nrow = niter + 1, ncol = 3)
+   x1 = matrix(0, nrow = niter - nburn, ncol = 3)
+   nv = array(0, 2)
+   nv1 = array(0, 2)
+   x[1, (1:3)] = c(1, 1, 1)
+   model = 1
+   t1 = prod(y)
+   t2 = sum(y)
+   for (i in 1:niter) {
+     if (model == 1) {
+       x[nv[1] + 1, 1] = rgamma(1, n + 2, t2 + 1)
+     }
+     else {
+       x[nv[2] + 1, 3] = rgamma(1, 4 + n * x[nv[2], 2],
+                                 t2 + 2)
+       x[nv[2] + 1, 2] = mh.alpha(y, n, x[nv[2], 2], x[nv[2] +
+                                                 1, 3], eps)
+     }
+     new = rj.modelo(y, n, x[nv[1] + 1, 1], x[nv[2] + 1, 2],
+                     x[nv[2] + 1, 3], model, a, b)
+     model = new$model
+     if (model == 1) {
+       x[nv[1] + 1, 1] = new$lambda
+       nv[1] = nv[1] + 1
+       if (i > nburn) {
+         x1[nv1[1] + 1, 1] = new$lambda
+       }
+     }
+   }
+ }
```

```

+
+           nv1[1] = nv1[1] + 1
+
+       }
+
+   else {
+
+       x[nv[2] + 1, 2] = new$alpha
+       x[nv[2] + 1, 3] = new$beta
+       nv[2] = nv[2] + 1
+       if (i > nburn) {
+
+           x1[nv1[2] + 1, 2] = new$alpha
+           x1[nv1[2] + 1, 3] = new$beta
+           nv1[2] = nv1[2] + 1
+
+       }
+
+   }
+
+   cat("Probabilidades a posteriori dos modelos", "\n")
+   print(nv1/(niter - nburn))
+   cat("Medias a posteriori dos parametros", "\n")
+   somas = apply(x1, 2, sum)
+   print(somas/c(nv1[1], nv1[2], nv1[2]))
+   return(list(x = x, nv = nv, x1 = x1, nv1 = nv1))
+
}

```

Vamos testar as funções acima simulando um conjunto de dados com distribuição exponencial.

```

> y = rexp(10, 3)
> niter = 1000
> nburn = 500
> m = rjmcmc(1000, 500, y, 10, 1, 1)

```

```

Probabilidades a posteriori dos modelos
[1] 0.8 0.2
Medias a posteriori dos parametros
[1] 3.794036 1.044988 3.439110

```

Assim o modelo exponencial tem probabilidade a posteriori bem maior que o modelo gama. Podemos estar interessados em estimar os tempos médios de vida ($E(Y)$) sob cada modelo.

```

> r1 = 1:m$nv1[1]
> r2 = 1:m$nv1[2]
> x = m$x1[, c(1, 2)]
> x[r1, 1] = 1/m$x1[r1, 1]

```

```

> x[r2, 2] = m$x1[r2, 2]/m$x1[r2, 3]
> somas = apply(x, 2, sum)
> medias = somas/c(m$nv1[1], m$nv1[2])
> print(medias)

[1] 0.2892936 0.3186531

> prob = m$nv1/(niter - nburn)
> prob[1] * medias[1] + prob[2] * medias[2]

[1] 0.2951655

```

4.8 Tópicos Relacionados

4.8.1 Autocorrelação Amostral

Em uma cadeia de Markov, os valores gerados são por definição correlacionados ao longo das iterações pois o valor de $\theta^{(t)}$ foi gerado a partir de $\theta^{(t-1)}$. Em muitas situações estes valores podem ser altamente correlacionados e em geral a autocorrelação será positiva. Ou seja, pode não haver muito ganho em termos de informação em se armazenar todos os valores simulados da cadeia e podemos estar desperdiçando espaço em disco, especialmente se a dimensão do problema for muito grande.

Embora não tenha nenhuma justificativa teórica, uma abordagem prática muito utilizada consiste em guardar os valores simulados a cada k iterações. Neste caso, dizemos que as simulações foram feitas com *thinning* igual a k . Por exemplo, se foram feitas 100 mil simulações, descartadas as 50 mil primeiras e guardados os valores a cada 10 iterações então no final as inferências serão baseadas em uma amostra de tamanho 5000.

Comentário

A não ser para obter esta redução de espaço ocupado em disco, descartar valores simulados (além daqueles da amostra de aquecimento) me parece um desperdício. Métodos de séries temporais estão disponíveis para analisar cadeias levando em conta as autocorrelações. Além disso pode-se tentar outros amostradores que gerem cadeias com menor autocorrelação amostral.

4.8.2 Monitorando a Convergência

Aqui vale lembrar que a verificação de convergência (ou falta de convergência) é responsabilidade do analista. Além disso estamos falando de convergência para

a distribuição alvo, que neste caso é a distribuição a posteriori, o que pode ser extremamente difícil de se verificar na prática.

Capítulo 5

Modelos Lineares

Em uma situação mais geral, a variável de interesse (variável resposta) tem sua descrição probabilística afetada por outras variáveis (variáveis explicativas ou covariáveis). No caso mais simples a influência sobre a resposta média é linear e aditiva e pode ser vista como uma aproximação de primeira ordem para funções mais complexas.

Usando uma notação matricial, o modelo linear normal pode ser escrito como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

onde \mathbf{y} é um vetor $n \times 1$ de observações, \mathbf{X} é uma matriz $n \times p$ conhecida, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de parâmetros e $\boldsymbol{\epsilon}$ é um vetor $n \times 1$ de erros aleatórios tais que $\epsilon_i \sim N(0, \sigma^2)$ e $E(\epsilon_i \epsilon_j) = 0$, para $i = 1, \dots, n$ e $j \neq i$. O modelo nos diz então que, a distribuição condicional de \mathbf{y} dados $\boldsymbol{\beta}$ e σ^2 é normal multivariada, i.e. $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ sendo \mathbf{I}_n é a matriz identidade de ordem n . Definindo $\phi = \sigma^{-2}$ e usando a função de densidade da normal multivariada (ver apêndice A) segue que

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \phi) &= (2\pi)^{-n/2} |\phi^{-1} \mathbf{I}_n|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\phi^{-1} \mathbf{I}_n)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &\propto \phi^{n/2} \exp \left\{ -\frac{\phi}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \end{aligned} \quad (5.1)$$

A forma quadrática em (5.1) pode ser reescrita em termos de $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ que é o estimador de máxima verossimilhança de $\boldsymbol{\beta}$,

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &\quad - 2(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \end{aligned}$$

pois $\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$. Denotando por $S = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ a soma de quadrados residual, podemos escrever então a função de verossimilhança como,

$$f(\mathbf{y}|\boldsymbol{\beta}, \phi) \propto \phi^{n/2} \exp \left\{ -\frac{\phi}{2} [(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + S] \right\}.$$

A distribuição a priori adotada aqui é uma generalização multivariada da distribuição Normal-Gama vista na Seção 2.3.5. Assim, a distribuição a priori é especificada como

$$\boldsymbol{\beta}|\phi \sim N_p(\boldsymbol{\mu}_0, (\mathbf{C}_0\phi)^{-1})$$

onde \mathbf{C}_0 é agora uma matriz $p \times p$ e

$$\phi \sim \text{Gama} \left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2} \right).$$

Com isso a densidade a priori conjunta de $(\boldsymbol{\beta}, \phi)$ fica completamente especificada e assim como no caso univariado a distribuição marginal de $\boldsymbol{\beta}$ é obtida integrando-se $p(\boldsymbol{\beta}, \phi)$ em relação a ϕ onde,

$$p(\boldsymbol{\beta}, \phi) \propto \phi^{\frac{n_0+p}{2}-1} \exp \left\{ -\frac{\phi}{2} [n_0\sigma_0^2 + (\boldsymbol{\beta}' - \boldsymbol{\mu}_0)' \mathbf{C}_0 (\boldsymbol{\beta}' - \boldsymbol{\mu}_0)] \right\}.$$

É fácil verificar que

$$p(\boldsymbol{\beta}) \propto \left[1 + \frac{(\boldsymbol{\beta} - \boldsymbol{\mu}_0)' \mathbf{C}_0 (\boldsymbol{\beta} - \boldsymbol{\mu}_0)}{n_0\sigma_0^2} \right]^{-(n_0+p)/2}$$

de modo que a distribuição a priori marginal de $\boldsymbol{\beta}$ é $\boldsymbol{\beta} \sim t_{n_0}(\boldsymbol{\mu}_0, \sigma_0^2 \mathbf{C}_0^{-1})$. Note que, como \mathbf{C}_0 é simétrica, é necessário especificar $p(p+1)/2$ de seus elementos. Na prática, podemos simplificar esta especificação assumindo que \mathbf{C}_0 é diagonal, i.e. que os componentes de $\boldsymbol{\beta}$ são não correlacionados a priori.

Combinando-se com a verossimilhança via teorema de Bayes obtem-se as seguintes distribuições a posteriori

$$\begin{aligned} \boldsymbol{\beta}|\phi, \mathbf{y} &\sim N(\boldsymbol{\mu}_1, (\mathbf{C}_1\phi)^{-1}) \\ \phi|\mathbf{y} &\sim \text{Gama} \left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2} \right) \quad \text{ou} \quad n_1\sigma_1^2\phi \sim \chi_{n_1}^2 \\ \boldsymbol{\beta}|\mathbf{y} &\sim t_{n_1}(\boldsymbol{\mu}_1, \sigma_1^2 \mathbf{C}_1^{-1}) \end{aligned}$$

onde os parâmetros atualizados são

$$\begin{aligned} n_1 &= n_0 + n \\ \mathbf{C}_1 &= \mathbf{C}_0 + \mathbf{X}'\mathbf{X} \\ \boldsymbol{\mu}_1 &= (\mathbf{C}_0 + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{C}_0\boldsymbol{\mu}_0 + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) \\ n_1\sigma_1^2 &= n_0\sigma_0^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}_1)'y + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'C_0\boldsymbol{\mu}_0 \\ &= n_0\sigma_0^2 + (n-p)\hat{\sigma}^2 + (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\beta}})'[\mathbf{C}_0^{-1} + \mathbf{X}'\mathbf{X}^{-1}]^{-1}(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\beta}}) \end{aligned}$$

onde

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Os estimadores pontuais de $\boldsymbol{\beta}$ e ϕ são dados respectivamente por $\boldsymbol{\mu}_1$ e σ_1^{-2} .

Intervalos de confiança para $\boldsymbol{\beta}_j$ e ϕ são obtidos através dos percentis das distribuições univariadas $t_{n_1}(\mu_j, \sigma_1^2(\mathbf{C}_1^{-1})_{jj})$, $j = 1, \dots, p$ e $\chi_{n_1}^2$. Em particular, note que $\boldsymbol{\mu}_1$ é obtida como uma ponderação matricial entre a estimativa a priori de $\boldsymbol{\beta}$ e sua estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$. Inferência conjunta sobre $\boldsymbol{\beta}$ também pode ser feita usando o fato que a forma quadrática

$$\frac{(\boldsymbol{\beta} - \boldsymbol{\mu}_1)'C_1(\boldsymbol{\beta} - \boldsymbol{\mu}_1)/p}{\sigma_1^2} \sim F(p, n_1).$$

Note que o modelo visto na seção anterior é na verdade o caso mais simples de um modelo linear quando $p = 1$ e \mathbf{X} é um vetor $n \times 1$ de 1's. Neste caso $\boldsymbol{\beta}$ é um escalar podendo ser denotado por μ e o modelo se reduz a $y_i = \mu + \epsilon_i$.

A priori não informativa é também uma generalização multivariada da seção anterior. Aqui o vetor $\boldsymbol{\beta}$ é um parâmetro de locação e ϕ é um parâmetro de escala, e portanto a priori não informativa de Jeffreys é $p(\boldsymbol{\beta}, \phi) \propto \phi^{-1}$. Vale notar que esta priori é um caso particular (degenerado) da priori conjugada natural com $\mathbf{C}_0 = 0$ e $n_0 = -p$. Fazendo as substituições adequadas obtém-se que as distribuições a posteriori são dadas por

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y} &\sim t_{n-p}(\hat{\boldsymbol{\beta}}, s^2(\mathbf{X}'\mathbf{X})^{-1}) \\ (n-p)s^2\phi|\mathbf{y} &\sim \chi_{n-p}^2 \\ \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'X'X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{s^2}|\mathbf{y} &\sim F(p, n-p) \end{aligned}$$

e estimadores pontuais bem como intervalos de confiança coincidirão com os obtidos usando métodos clássicos.

5.1 Análise de Variância com 1 Fator de Classificação

Considere o modelo $y_{ij} = \beta_j + \epsilon_{ij}$, $i = 1, \dots, n_j$ e $j = 1, \dots, p$. Assim, todas as n_j observações do grupo j têm a mesma média β_j . Neste problema, o número total de observações independentes é $n = n_1 + \dots + n_p$. Em outras palavras, $Y_{1j}, \dots, Y_{n_j j} \sim N(\beta_j, \sigma^2)$. Se os y_{ij} forem “empilhados” em um único vetor $n \times 1$ então podemos reescrever o modelo na forma matricial $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ sendo

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Note que $\mathbf{X}'\mathbf{X} = \text{diagonal}(n_1, \dots, n_p)$ e a forma quadrática $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ se reduz a

$$\sum_{j=1}^p n_j (\beta_j - \bar{y}_j)^2$$

e a função de verossimilhança é dada por

$$l(\beta_1, \dots, \beta_p, \phi; y) \propto \phi^{n/2} \exp \left\{ -\frac{\phi}{2} \left[(n-p)s^2 + \sum_{j=1}^p n_j (\beta_j - \bar{y}_j)^2 \right] \right\}$$

com

$$s^2 = \frac{1}{n-p} (y - X\hat{\beta})'(y - X\hat{\beta}).$$

Assumindo que $\beta_j | \phi \sim N(\mu_j, (c_j \phi)^{-1})$, $j = 1, \dots, p$ são condicionalmente independentes e que $n_0 \sigma_0^2 \phi \sim \chi_{n_0}^2$ então as distribuições a posteriori são

$$\begin{aligned} \beta_j | \phi, y &\sim N(\mu_j^*, (c_j^* \phi)^{-1}) \\ n_1 \sigma_1^2 \phi | y &\sim \chi_{n_1}^2 \\ \beta_j | y &\sim t_{n_1}(\mu_j^*, \sigma_1^2 / c_j^*) \end{aligned}$$

onde

$$\begin{aligned}\mu_j^* &= \frac{c_j\mu_j + n_j\bar{y}_j}{c_j + n_j} \\ c_j^* &= c_j + n_j \\ n_1 &= n_0 + n \\ n_1\sigma_1^2 &= n_0\sigma_0^2 + (n-p)s^2 + \sum_{i=1}^p \frac{n_j c_j}{c_j + n_j} (\bar{y}_j - \mu_j)^2\end{aligned}$$

e os $\beta_j|\phi, y$ permanecem independentes.

A priori não informativa $p(\beta, \phi) \propto \phi^{-1}$ é obtida fazendo-se $c_j = 0$, $j = 1, \dots, p$ e $n_0 = -p$. Assim, as distribuições a posteriori marginais são dadas por

$$\begin{aligned}\beta_j|y &\sim t_{n-p}(\bar{y}_j, s^2/n_j) \\ (n-p)s^2\phi &\sim \chi_{n-p}^2\end{aligned}$$

e as estimativas pontuais e intervalos de confiança coincidirão com os da inferência clássica. Em particular, se estamos interessados em testar

$$H_0 : \beta_1 = \dots = \beta_p = \beta$$

então pode-se mostrar que (DeGroot, 1970, páginas 257 a 259) deve-se rejeitar H_0 se

$$P\left(F > \frac{\sum_{j=1}^p n_j (\bar{y}_j - \bar{\bar{y}})^2 / (p-1)}{s^2}\right)$$

onde $F \sim F(p-1, n-p)$ for pequena.

Note que as hipóteses equivalentes são

$$H_0 : \alpha_1 = \dots = \alpha_p = 0$$

sendo

$$\alpha_j = \beta_j - \beta, \quad \beta = \frac{1}{n} \sum_{j=1}^p n_j \beta_j \quad \text{e} \quad \sum_{j=1}^p n_j \alpha_j = 0$$

e α_j é o efeito da j -ésima população. Neste caso, $X'X = \text{diagonal}(n_1, \dots, n_p)$ e a forma quadrática $(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})$ fica $\sum n_j(\alpha_j - \bar{y}_j - \bar{\bar{y}})^2 + n(\beta - \bar{y}_j - \bar{\bar{y}})^2$.

Apêndice A

Listas de Distribuições

Neste apêndice são listadas as distribuições de probabilidade utilizadas no texto para facilidade de referência. Só apresentadas suas funções de (densidade) de probabilidade além da média e variância. Uma revisão exaustiva de distribuições de probabilidades pode ser encontrada em Johnson et al. (1992, 1995) e Evans et al. (1993).

A.1 Distribuição Normal

X tem distribuição normal com parâmetros $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, denotando-se $X \sim N(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right], \quad -\infty < x < \infty.$$

$$E(X) = \mu \quad \text{e} \quad V(X) = \sigma^2.$$

Quando $\mu = 0$ e $\sigma^2 = 1$ a distribuição é chamada normal padrão.

No caso vetorial, $\mathbf{X} = (X_1, \dots, X_p)$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de variância-covariância Σ , denotando-se $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ se sua função de densidade é dada por

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp[-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})/2]$$

para $\boldsymbol{\mu} \in \mathbb{R}^p$ e Σ positiva-definida.

A.2 Distribuição Log-Normal

Se $X \sim N(\mu, \sigma^2)$ então $Y = e^X$ tem distribuição log-normal com parâmetros μ e σ^2 . Portanto, sua função de densidade é dada por

$$p(y|\mu, \sigma^2) = \frac{1}{y} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2}\frac{(\log(y) - \mu)^2}{\sigma^2}\right], \quad -\infty < x < \infty.$$

$$E(X) = \exp\{\mu + \sigma^2/2\} \quad \text{e} \quad V(X) = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1).$$

A.3 A Função Gama

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

Propriedades,

- Usando integração por partes pode-se mostrar que,

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \alpha > 0.$$

- $\Gamma(1) = 1$.
- $\Gamma(1/2) = \sqrt{\pi}$.
- Para n um inteiro positivo,

$$\Gamma(n+1) = n! \quad \text{e} \quad \Gamma\left(n + \frac{1}{2}\right) = \left(n - \frac{1}{2}\right)\left(n - \frac{3}{2}\right) \cdots \frac{3}{2} \frac{1}{2} \sqrt{\pi}$$

A.4 Distribuição Gama

X tem distribuição Gama com parâmetros $\alpha > 0$ e $\beta > 0$, denotando-se $X \sim Ga(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0.$$

$$E(X) = \alpha/\beta \quad \text{e} \quad V(X) = \alpha/\beta^2.$$

Casos particulares da distribuição Gama são a distribuição de Erlang, $Ga(\alpha, 1)$, a distribuição exponencial, $Ga(1, \beta)$, e a distribuição qui-quadrado com ν graus de liberdade, $Ga(\nu/2, 1/2)$.

A.5 Distribuição Wishart

Diz-se que uma matriz aleatória Ω ($n \times n$) segue uma distribuição Wishart com parâmetro Σ e ν graus de liberdade, denotando-se $\Omega \sim W(\Sigma, \nu)$, se sua função de densidade é dada por,

$$p(\Omega|\Sigma, \nu) \propto |\Omega|^{(\nu-n-1)/2} \exp(-(1/2)\text{tr}(\Sigma\Omega))$$

sendo $\nu \geq n$, Σ positiva-definida e $\text{tr}(A)$ indica o traço de uma matriz A . Uma propriedade útil é que $A\Omega A' \sim W(A\Sigma A', \nu)$.

A.6 Distribuição Gama Inversa

X tem distribuição Gama Inversa com parâmetros $\alpha > 0$ e $\beta > 0$, denotando-se $X \sim GI(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}, \quad x > 0.$$

$$E(X) = \frac{\beta}{\alpha - 1}, \quad \text{para } \alpha > 1 \quad \text{e} \quad V(X) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}, \quad \text{para } \alpha > 2.$$

Não é difícil verificar que esta é a distribuição de $1/X$ quando $X \sim Ga(\alpha, \beta)$.

A.7 Distribuição Wishart Invertida

Diz-se que uma matriz aleatória Ω ($n \times n$) segue uma distribuição Wishart-Invertida com parâmetro Σ e ν graus de liberdade, denotando-se $\Omega \sim WI(\Sigma, \nu)$ se sua função de densidade é dada por,

$$p(\Omega|\Sigma, \nu) \propto |\Omega|^{-(\nu+n+1)/2} \exp(-(1/2)\text{tr}(\Sigma\Omega))$$

sendo $\nu \geq n$, Σ positiva-definida e $\text{tr}(A)$ indica o traço de uma matriz A . Não é difícil verificar que $\Omega^{-1} \sim W(\Sigma, \nu)$. Outra propriedade é que $A\Omega A' \sim WI(A\Sigma A', \nu)$.

A.8 Distribuição Beta

X tem distribuição Beta com parâmetros $\alpha > 0$ e $\beta > 0$, denotando-se $X \sim Be(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 < x < 1.$$

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

A.9 Distribuição de Dirichlet

O vetor aleatório $\mathbf{X} = (X_1, \dots, X_k)$ tem distribuição de Dirichlet com parâmetros $\alpha_1, \dots, \alpha_k$, denotada por $D_k(\alpha_1, \dots, \alpha_k)$ se sua função de densidade conjunta é dada por

$$p(\mathbf{x}|\alpha_1, \dots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \dots, \Gamma(\alpha_k)} x_1^{\alpha_1-1} \dots x_k^{\alpha_k-1}, \quad \sum_{i=1}^k x_i = 1,$$

para $\alpha_1, \dots, \alpha_k > 0$ e $\alpha_0 = \sum_{i=1}^k \alpha_i$.

$$E(X_i) = \frac{\alpha_i}{\alpha_0}, \quad V(X_i) = \frac{(\alpha_0 - \alpha_i)\alpha_i}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{e} \quad Cov(X_i, X_j) = -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Note que a distribuição Beta é obtida como caso particular para $k = 2$.

A.10 Distribuição t de Student

X tem distribuição t de Student (ou simplesmente t) com parâmetros $\mu \in \mathbb{R}$, $\sigma^2 > 0$ e $\nu > 0$ (chamado graus de liberdade), denotando-se $X \sim t_\nu(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi}\sigma} \left[\nu + \frac{(x-\mu)^2}{\sigma^2} \right]^{-(\nu+1)/2}, \quad x \in \mathbb{R}.$$

$$E(X) = \mu, \quad \text{para } \nu > 1 \quad \text{e} \quad V(X) = \sigma^2 \frac{\nu}{\nu - 2}, \quad \text{para } \nu > 2.$$

Um caso particular da distribuição t é a distribuição de Cauchy, denotada por $C(\mu, \sigma^2)$, que corresponde a $\nu = 1$.

A.11 Distribuição F de Fisher

X tem distribuição F com $\nu_1 > 0$ e $\nu_2 > 0$ graus de liberdade, denotando-se $X \sim F(\nu_1, \nu_2)$, se sua função de densidade é dada por

$$p(x|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1 + \nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{\nu_1/2-1} (\nu_2 + \nu_1 x)^{-(\nu_1 + \nu_2)/2}, \quad x > 0.$$

$$E(X) = \frac{\nu_2}{\nu_2 - 2}, \quad \text{para } \nu_2 > 2 \quad \text{e} \quad V(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)(\nu_2 - 2)^2}, \quad \text{para } \nu_2 > 4.$$

A.12 Distribuição de Pareto

X tem distribuição de Pareto com parâmetros α e β denotando-se $X \sim \text{Pareto}(\alpha, \beta)$, se sua função de densidade de probabilidade é dada por,

$$p(x|\alpha, \beta) = \frac{\alpha}{\beta} \left(\frac{\beta}{x} \right)^{\alpha+1}, \quad x > \beta.$$

$$E(X) = \frac{\alpha\beta}{\alpha - 1} \quad \text{e} \quad V(X) = \frac{\alpha\beta^2}{\alpha - 2} - \left(\frac{\alpha\beta}{\alpha - 1} \right)^2.$$

A.13 Distribuição Binomial

X tem distribuição binomial com parâmetros $n \geq 1$ e $p \in (0, 1)$, denotando-se $X \sim \text{bin}(n, p)$, se sua função de probabilidade é dada por

$$p(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \dots, n.$$

$$E(X) = np \quad \text{e} \quad V(X) = np(1-p)$$

e um caso particular é a distribuição de Bernoulli com $n = 1$.

A.14 Distribuição Multinomial

O vetor aleatório $\mathbf{X} = (X_1, \dots, X_k)$ tem distribuição multinomial com parâmetros n e probabilidades $\theta_1, \dots, \theta_k$, denotada por $M_k(n, \theta_1, \dots, \theta_k)$ se sua função de probabilidade conjunta é dada por

$$p(\mathbf{x}|\theta_1, \dots, \theta_k) = \frac{n!}{x_1! \dots x_k!} \theta_1^{x_1} \dots \theta_k^{x_k}, \quad x_i = 0, \dots, n, \quad \sum_{i=1}^k x_i = n,$$

para $0 < \theta_i < 1$ e $\sum_{i=1}^k \theta_i = 1$. Note que a distribuição binomial é um caso particular da distribuição multinomial quando $k = 2$. Além disso, a distribuição marginal de cada X_i é binomial com parâmetros n e θ_i e

$$E(X_i) = n\theta_i, \quad V(X_i) = n\theta_i(1 - \theta_i), \quad \text{e} \quad Cov(X_i, X_j) = -n\theta_i\theta_j.$$

A.15 Distribuição de Poisson

X tem distribuição de Poisson com parâmetro $\theta > 0$, denotando-se $X \sim Poisson(\theta)$, se sua função de probabilidade é dada por

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \dots$$

$$E(X) = V(X) = \theta.$$

A.16 Distribuição Binomial Negativa

X tem distribuição de binomial negativa com parâmetros $r \geq 1$ e $p \in (0, 1)$, denotando-se $X \sim BN(r, p)$, se sua função de probabilidade é dada por

$$p(x|r, p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, \dots$$

$$E(X) = \frac{r(1-p)}{p} \quad \text{e} \quad V(X) = \frac{r(1-p)}{p^2}.$$

Um caso particular é quando $r = 1$ e neste caso diz-se que X tem distribuição geométrica com parâmetro p . Neste caso,

$$p(x|p) = p^r (1-p)^x, \quad x = 0, 1, \dots$$

$$E(X) = \frac{1-p}{p} \quad \text{e} \quad V(X) = \frac{1-p}{p^2}.$$

Apêndice B

Alguns Endereços Interessantes

Neste apêndice são listados alguns endereços na internet com conteúdo relativo a abordagem Bayesiana.

- Teorema de Bayes no Wikipedia: http://en.wikipedia.org/wiki/Bayes_theorem
- Bayesian Analysis - The Journal: <http://ba.stat.cmu.edu/>
- International Society for Bayesian Analysis: <http://www.bayesian.org>
- American Statistical Association, Section on Bayesian Statistical Science: <http://www.amstat.org/sections/SBSS>
- Bayes Methods Working Group of the International Biometric Society, German Region: <http://ibealt.web.med.uni-muenchen.de/bayes-ag>
- Encontro Brasileiro de Estatística Bayesiana:
2006 (<http://www.im.ufrj.br/ebeb8>),
2008 (<http://www.ime.usp.br/~isbra/ebeb/9ebeb>)
- Valencia Meetings: <http://www.uv.es/valenciameting>
- I Workshop em Estatística Espacial e Métodos Computacionalmente Intensivos: leg.ufpr.br/~ehlers/folder
- Case Studies in Bayesian Statistics: <http://lib.stat.cmu.edu/bayesworkshop/>
- MCMC preprints: <http://www.statslab.cam.ac.uk/~mcmc>
- Projeto BUGS (Bayesian inference Using Gibbs Sampling): <http://www.mrc-bsu.cam.ac.uk/bugs>
- Projeto JAGS (Just Another Gibbs Sampler): <http://www-fis.iarc.fr/~martyn/software/jags/>

- BayesX (Bayesian Inference in Structured Additive Regression Models.):
<http://www.stat.uni-muenchen.de/~bayesx/bayesx.html>
- MrBayes (Bayesian estimation of phylogeny): <http://mrbayes.scs.fsu.edu>
- Número especial do Rnews dedicado a inferencia Bayesiana e MCMC:
http://www.est.ufpr.br/R/doc/Rnews/Rnews_2006-1.pdf
- CRAN Task View (Bayesian Inference):
<http://cran.r-project.org/src/contrib/Views/Bayesian.html>
- Centro de Estudos do Risco UFSCAR:
<http://www.ufscar.br/~des/CER/inicial.htm>

Referências

- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York.
- Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. Wiley: New York.
- Box, G. E. P. & Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley Classics Library ed. Wiley-Interscience.
- Broemeling, L. (1985). *Bayesian Analysis of Linear Models*. New York: Marcel Dekker.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill Book Co.
- Evans, M., Hastings, N. & Peacock, B. (1993). *Statistical Distributions, Second Edition* (Second ed.). Wiley Interscience.
- Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Sciences. Chapman and Hall, London.
- Gamerman, D. & Lopes, H. (2006). *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science Series. CRC Press.
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman and Hall: London.
- Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions* (2nd ed.), Volume 2. John Wiley, New York.
- Johnson, N. L., Kotz, S. & Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd ed.). John Wiley, New York.
- Migon, H. S. & Gamerman, D. (1999). *Statistical Inference: An Integrated Approach*. Arnold.
- O'Hagan, A. (1994). *Bayesian Inference*, Volume 2B. Edward Arnold, Cambridge.

- Robert, C. P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.
- Smith, A. F. M. & Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician* 46, 84–88.