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.

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

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

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

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 $p(\theta|\boldsymbol{x}) = kq(\theta)$, segue que
$$k = \left[\int q(\theta)d\theta\right]^{-1}.$$
2. Se $g(\theta) = \theta$, então têm-se $\mu = E(\theta|\boldsymbol{x})$, média a posteriori.
3. Quando $g(\theta) = (\theta - \mu)^2$, então $\sigma^2 = E((\theta - \mu)^2|\boldsymbol{x})$, 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 $P(A \mid \boldsymbol{x}) = \int_A p(\theta|\boldsymbol{x})d\theta$
5. Seja $g(\theta) = p(y|\theta)$, onde $y \perp \boldsymbol{x}|\theta$. Nestas condições obtemos $E[p(y|\boldsymbol{x})]$, 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, $\theta_1, \ldots, \theta_n$ da distribuição uniforme no intervalo (a, b) teremos também uma amostra de valores $g(\theta_1), \ldots, g(\theta_n)$ 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 viesada 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 $\theta_1, \ldots, \theta_n$ da distribuição $U(a, b)$;
2. calcule $g(\theta_1), \ldots, g(\theta_n)$;
3. calcule a média amostral $\overline{g} = \sum_{i=1}^{n} g(\theta_i)/n$
4. calcule $\hat{I} = (b-a)\overline{g}$

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 $y_i = e^{-x_i}$, $i = 1, \ldots, 100$. 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.

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.

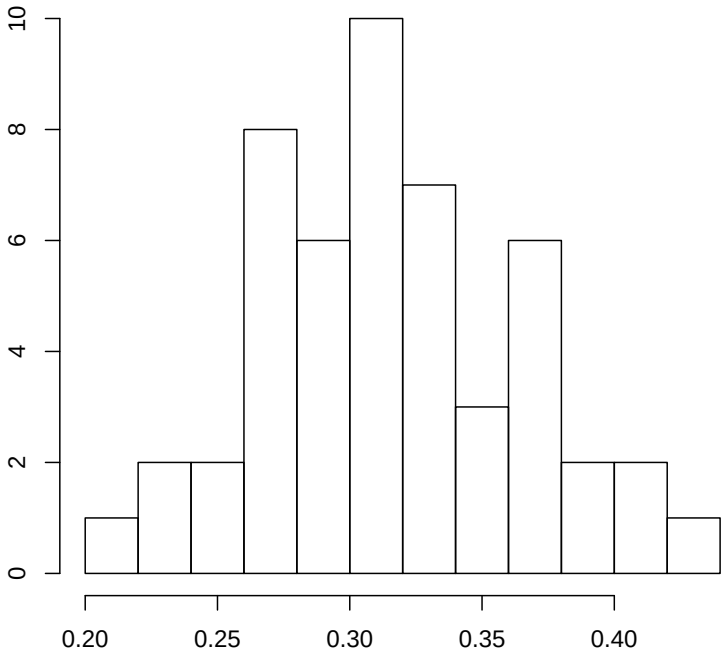

Figura 4.1: Histograma de 50 estimativas de Monte Carlo da integral no Exemplo 4.1 com $n = 10$.

A generalização é bem simples para o caso em que a integral é a esperança

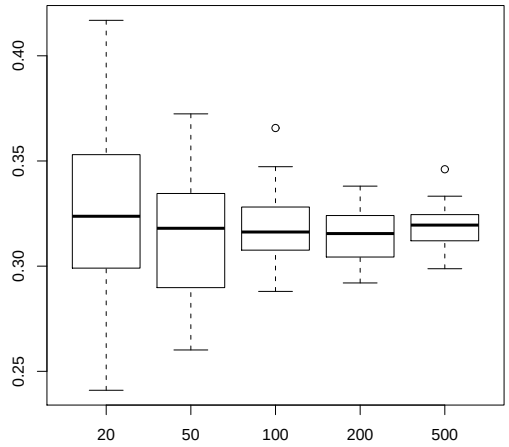

Figura 4.2: Boxplots para 50 estimativas da integral no Exemplo 4.1 com n=20, 50, 100, 200, e 500 simulações.

matemática de uma função $g(\theta)$ onde θ tem função de densidade $p(\theta)$, i.e.

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

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

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

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

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

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

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

e portanto a variância do estimador pode também ser estimada como

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

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 grande segue que

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

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ um vetor aleatório de dimensão k com função de densidade $p(\boldsymbol{\theta})$. Neste caso os valores gerados serão também vetores $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ 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 camandos do `R` abaixo.

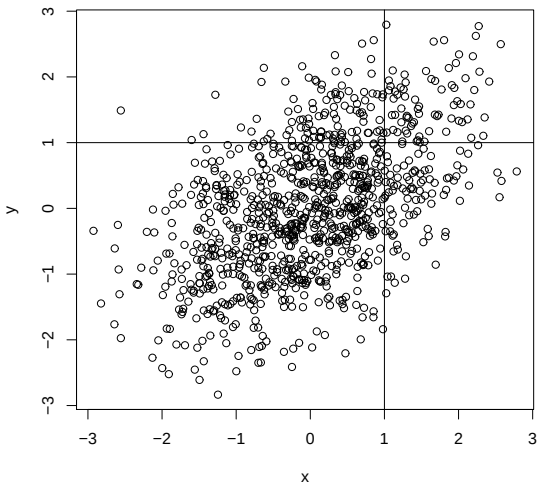

Figura 4.3: Diagrama de dispersão de 1000 valores simulados da distribuição N(0,1) bivariada.

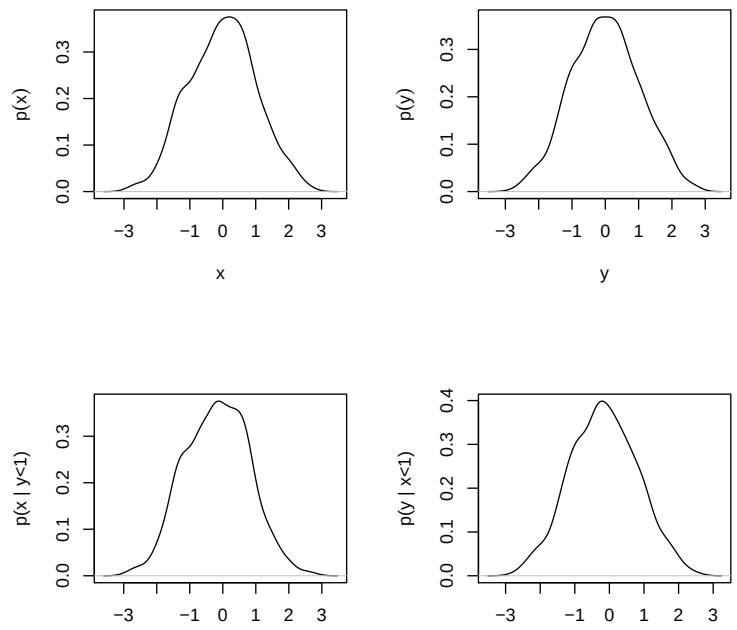

Figura 4.4: Estimativas das densidades marginais e condicionais no Exemplo 4.2.

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*.

Se $q(\theta)$ for uma função de densidade definida no mesmo espaço de 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 $\theta_1, \ldots, \theta_n$ 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

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

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

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

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 $\theta_1, \ldots, \theta_n$ independentes da distribuição $N(x, 1)$;

2. calcular $g_i = \dfrac{\theta_i}{1+\theta_i^2}$ e $g_i^* = \dfrac{1}{1+\theta_i^2}$;

3. calcular $\hat{E}(\theta|\boldsymbol{x}) = \dfrac{\sum_{i=1}^{n} g_i}{\sum_{i=1}^{n} g_i^*}$.

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.

Exercicios

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 superimposta. 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 $X_1, \ldots, X_k$ 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 $p(\theta|\boldsymbol{x}) < Aq(\theta)$. 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

$p(\theta^*|\boldsymbol{x})/Aq(\theta^*)$. 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 $p(\boldsymbol{x}|\theta)p(\theta)$.

Podemos então usar o seguinte algoritmo,

1. gerar um valor θ^* da distribuição auxiliar;
2. gerar $u \sim U(0,1)$;
3. se $u < p(\theta^*|\boldsymbol{x})/Aq(\theta^*)$ 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 $p(\boldsymbol{x}|\theta) < A$. Esta desigualdade é satisfeita se tomarmos A como sendo o valor máximo da função de verossimilhança, i.e. $A = p(\boldsymbol{x}|\hat{\theta})$ onde $\hat{\theta}$ é o estimador de máxima verossimilhança de θ. Neste caso, a probabilidade de aceitação se simplifica para $p(\boldsymbol{x}|\theta)/p(\boldsymbol{x}|\hat{\theta})$.

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 $u < p(\boldsymbol{x}|\theta^*)/p(\boldsymbol{x}|\hat{\theta})$, caso contrário rejeitar θ^* e retornar ao passo 1.

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

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

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.

```
Taxa de aceitacao 0.022
```

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.

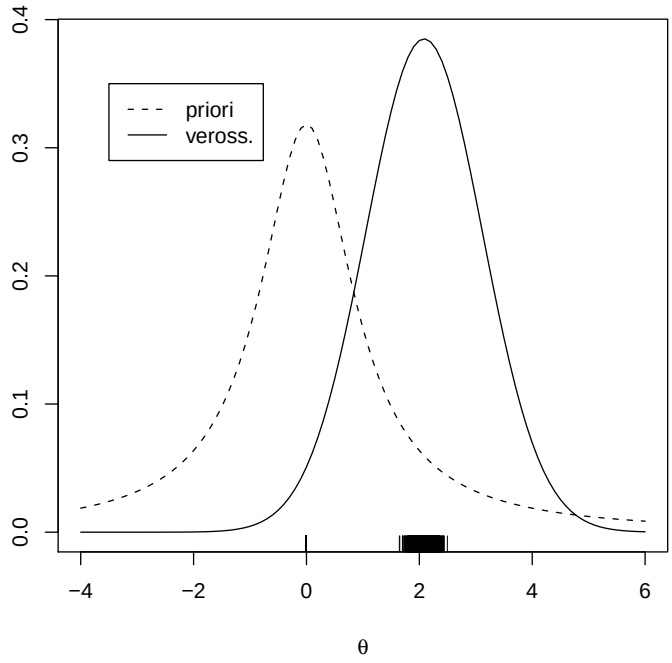

Figura 4.5: Verossimilhança normalizada e densidade a priori juntamente com valores simulados.

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.

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 $\theta_1, \ldots, \theta_n$ gerada da distribuição auxiliar q e a partir dela construimos os pesos

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

O método consiste em tomar uma segunda amostra (ou reamostra) de tamanho m da distribuição discreta em $\theta_1, \ldots, \theta_n$ com probabilidades $w_1, \ldots, w_n$. Aqui também não é necessário que se conheça completamente a densidade 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

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

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

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

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.

Exercícios

1. Em um modelo de regressão linear simples temos que $y_i \sim N(\beta x_i, 1)$. Os dados observados são $\boldsymbol{y} = (-2, 0, 0, 0, 2)$ e $\boldsymbol{x} = (-2, -1, 0, 1, 2)$, e usamos uma priori vaga $N(0, 4)$ para β. Faça inferência sobre β obtendo uma

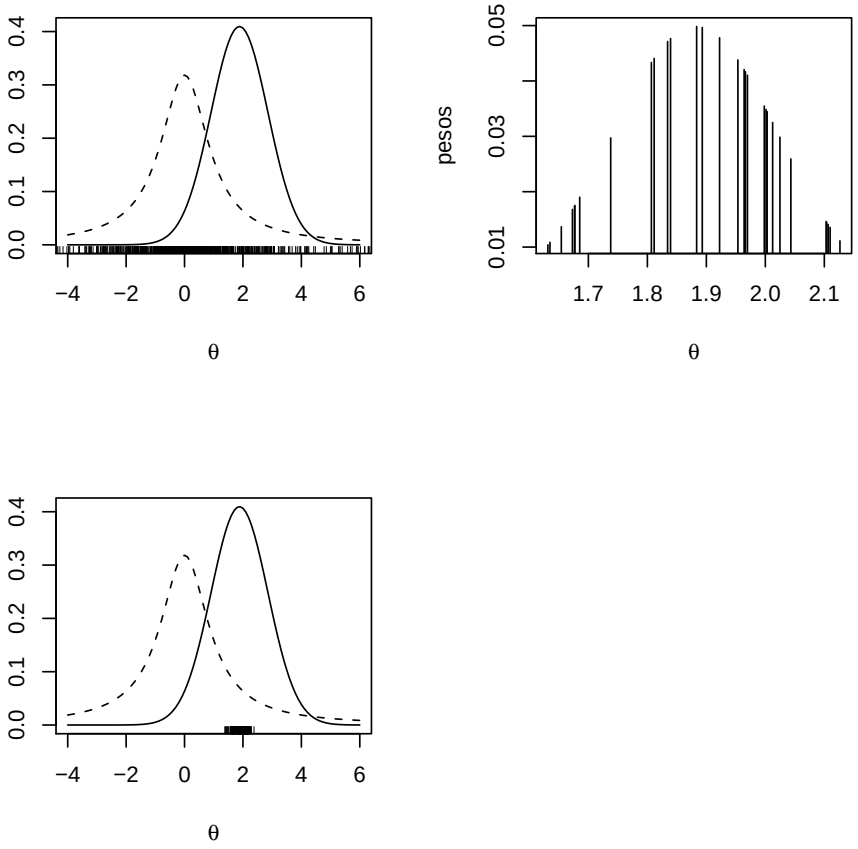

Figura 4.6: Verossimilhança normalizada (linha cheia), densidade a priori (linha tracejada) e os valores amostrados (a) e reamostrados (c). Em (b) os valores de θ com pesos maiores do que 0,01.

amostra da posteriori usando reamostragem ponderada. Compare com a estimativa de máxima verossimilhança $\hat{\beta} = 0,8$.

2. 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 $\sigma^{-2} \sim G(0,01, 0,01)$.

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

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

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 tomando

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

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

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

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 equilibrio, 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

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

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,

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

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

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

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 $\boldsymbol{Y} = (y_1, y_2, y_3, y_4)$ tem distribuição multinomial com parâmetros n, p_1, p_2, p_3, p_4 e portanto,

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

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

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

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.

Suponha que foram observados 197 animais com os números de animais nas categorias dados por $\boldsymbol{y} = (125, 18, 20, 34)$ e foi gerada uma cadeia de Markov com 10000 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 tambem 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.

```
[1] 0.1639
```

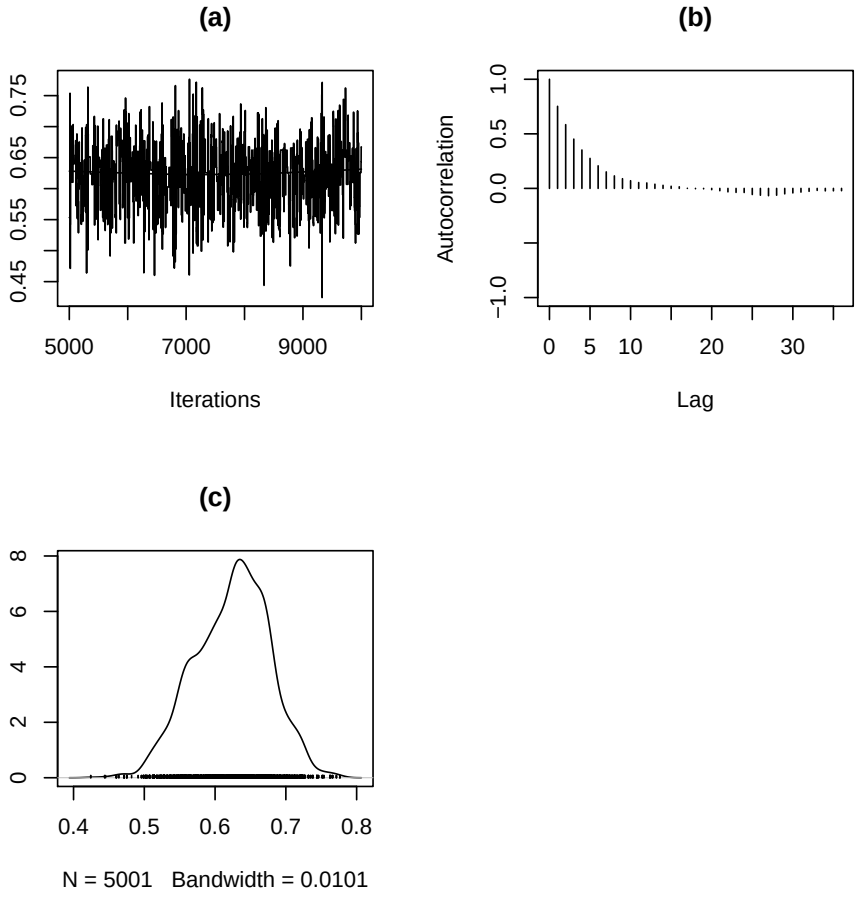

Figura 4.7: (a) 10000 valores simulados de θ, (b) 30 primeiras autocorrelações amostrais após aquecimento, (c) Densidade a posteriori estimada.

Dada uma amostra com valores de θ temos também amostras de valores de (p_1, p_2, p_3, p_4) que estão resumidos na Tabela 4.1

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 $\sigma = 10$.

	Mean	SD	Naive SE	Time-series SE	2.5%	25%	50%	75%	97.5%
p1	0.656	0.013	0.000	0.000	0.629	0.647	0.657	0.665	0.680
p2	0.094	0.013	0.000	0.000	0.070	0.085	0.093	0.103	0.121
p3	0.094	0.013	0.000	0.000	0.070	0.085	0.093	0.103	0.121
p4	0.156	0.013	0.000	0.000	0.129	0.147	0.157	0.165	0.180

Tabela 4.1:

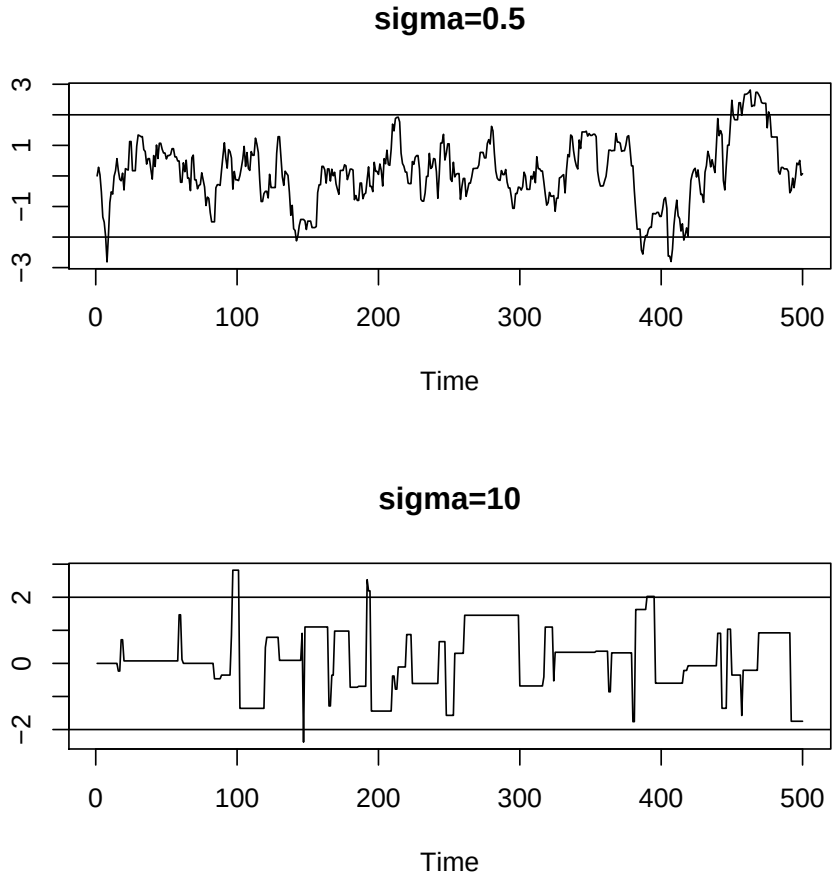

Figura 4.8: 500 valores simulados para o Exemplo 4.7 usando o algoritmo de Metropolis-Hastings com (a) $\sigma = 0.5$ e (b) $\sigma = 10$.

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,

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

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

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

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

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

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

Exercicios

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 $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)'$.

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,

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

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 $\boldsymbol{y}$ que depende dos parâmetros θ, λ e δ suponha que a distribuição conjunta é dada por

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

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

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

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 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})'$;
3. obtenha um novo valor de $\boldsymbol{\theta}^{(t)}$ a partir de $\boldsymbol{\theta}^{(t-1)}$ através da geração sucessiva dos valores

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

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

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

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

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

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

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

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

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

Segue então que

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

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

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

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

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

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

a densidade a posteriori fica

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

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

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

O algoritmo foi testado com 40 dados simulados de processos com médias 2 e 5 e ponto de mudança 23. Foram rodadas 10000 iterações. As 5000 primeiras simulações foram descartadas como amostra de aquecimento.

	Mean	SD	Naive SE	Time-series SE	2.5%	25%	50%	75%	97.5%
lambda	2.234	0.428	0.006	0.012	1.491	1.999	2.229	2.479	3.043
phi	4.018	0.625	0.009	0.018	2.933	3.672	4.024	4.399	5.112
m	21.647	4.825	0.068	0.155	2.000	21.000	22.000	23.000	29.000

Tabela 4.2:

A partir dos valores simulados de m podemos estimar suas probabilidades a posteriori (Tabela 4.10). Finalmente, pode-se estimar as contagens médias condicionando no valor de m com maior probabilidade (Figura 4.11).

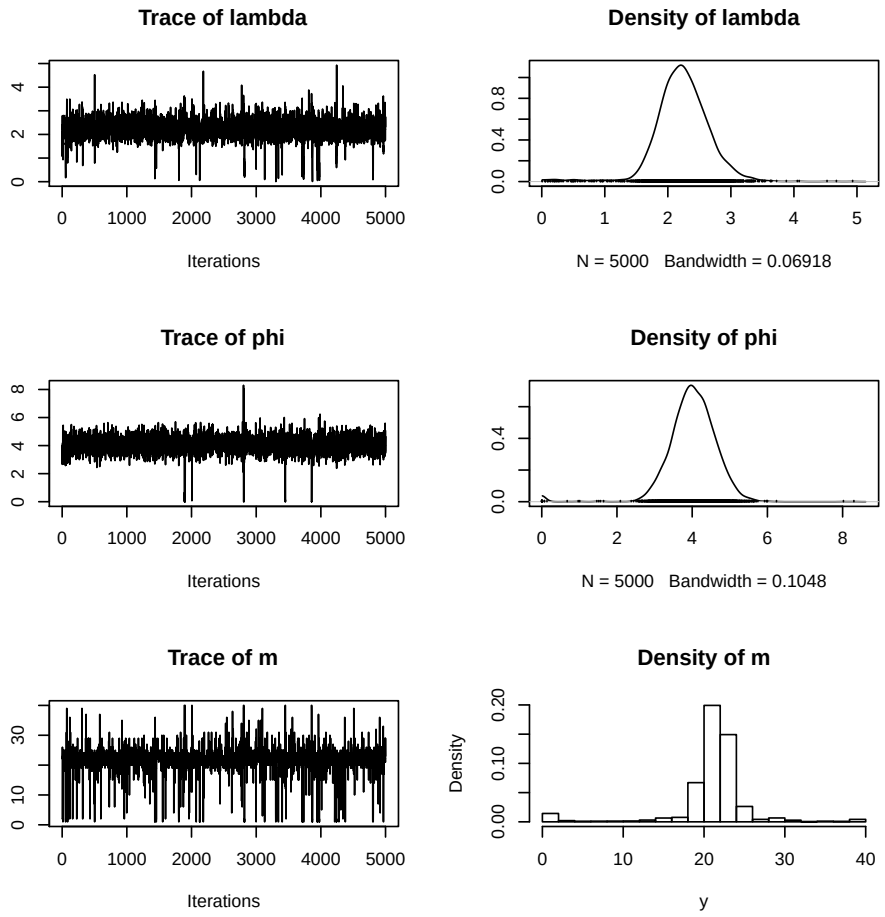

Figura 4.9:

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

Se M é conjunto de todos os possíveis modelos (ou modelos candidatos), então

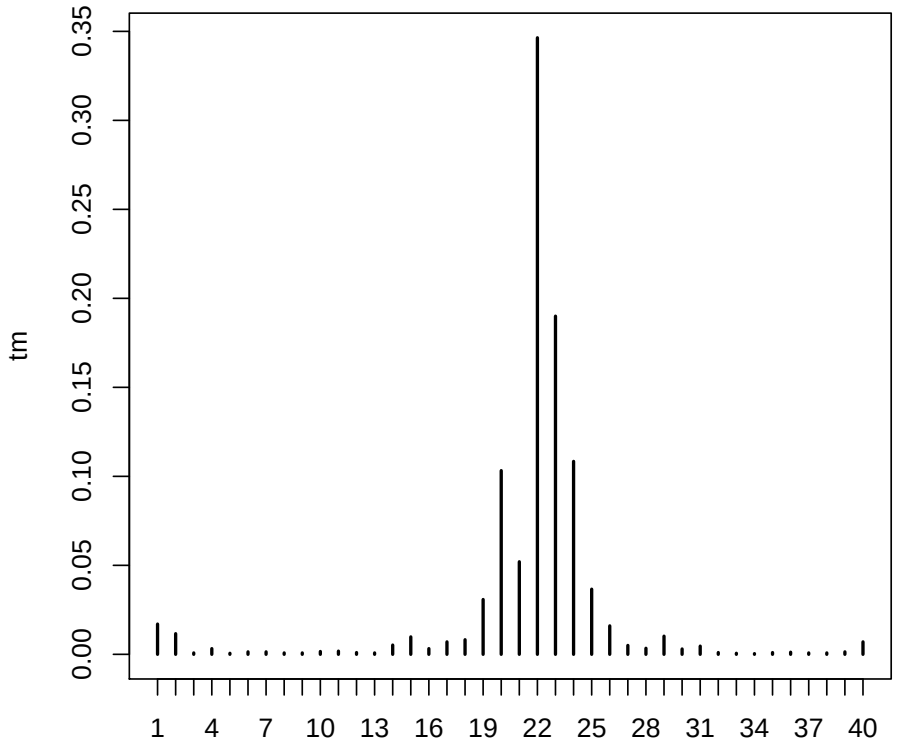

Figura 4.10:

as probabilidades a posteriori de cada possível modelo são dadas por

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

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

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

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) $p(\boldsymbol{y}|k)$ pode ser inviável na prática.

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

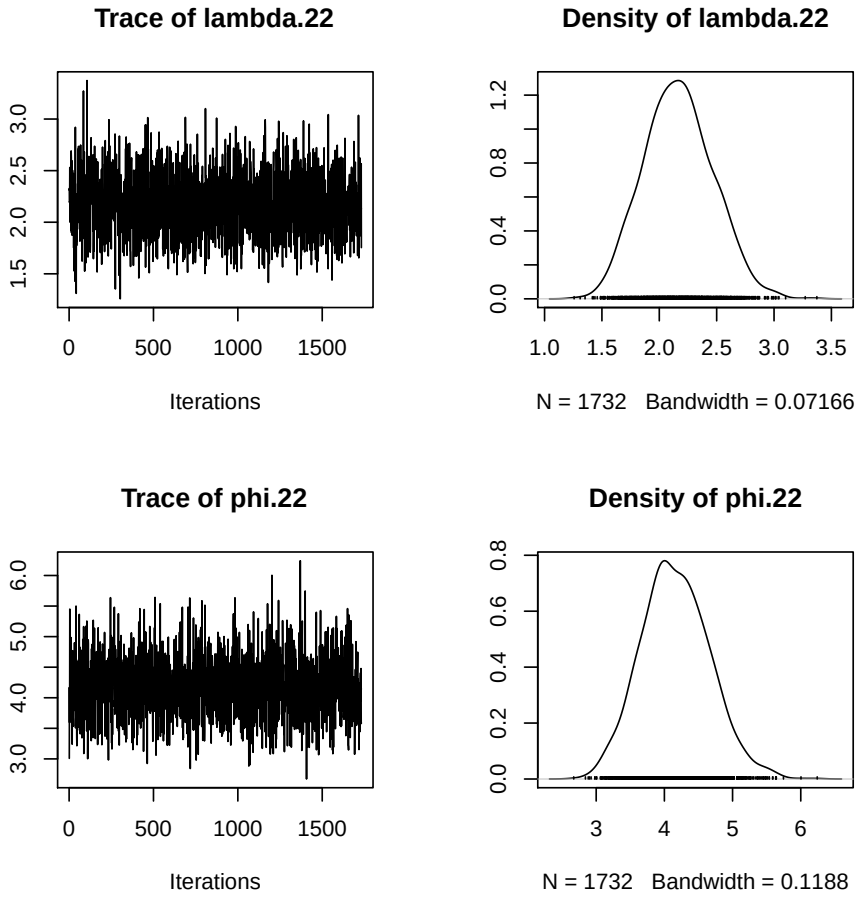

Figura 4.11:

posteriori conjunta,

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

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

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 reversiveis (Green 1995) que são particularmente úteis no contexto de seleção Bayesiana de modelos.

4.7.1 MCMC com Saltos Reversiveis (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 $\boldsymbol{\theta}' = g(\boldsymbol{\theta}, \boldsymbol{u})$ para uma função deterministica g e um vetor aleatório $\boldsymbol{u} \sim q(\boldsymbol{u})$ com dimensão $n_{k'} - n_k$. Então o seguinte algoritmo é utilizado,

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

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

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

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

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

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

Dado o modelo, as funções de verossimilhança ficam

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

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

as distribuições condicionais completas são facilmente obtidas como

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

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.

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

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

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 $u=\alpha$.

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

```
Probabilidades a posteriori dos modelos
[1] 0.796 0.204
Medias a posteriori dos parametros
[1] 3.557093 0.933466 3.074695
```

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 $\boldsymbol{\theta}^{(t)}$ foi gerado a partir de $\boldsymbol{\theta}^{(t-1)}$. 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 desperdicio. Métodos de séries temporais estão disponiveis 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.

4.9 Normalidade assintótica

Se a distribuição a posteriori for unimodal e aproximadamente simétrica ela pode ser aproximada por uma distribuição normal centrada na moda a posteriori. Considere a expansão de Taylor de $\log p(\theta|\boldsymbol{x})$ em torno da moda θ^*,

$$\begin{aligned}\log p(\theta|\boldsymbol{x}) = \log p(\theta^*|\boldsymbol{x}) &+ (\theta-\theta^*)\left[\frac{d}{d\theta}\log p(\theta|\boldsymbol{x})\right]_{\theta=\theta^*} \\ &+ \frac{1}{2}(\theta-\theta^*)^2\left[\frac{d^2}{d\theta^2}\log p(\theta|\boldsymbol{x})\right]_{\theta=\theta^*} + \ldots\end{aligned}$$

Por definição, $\left[\frac{d}{d\theta}\log p(\theta|\boldsymbol{x})\right]_{\theta=\theta^*} = 0$ e definindo $h(\theta) = \left[\frac{d^2}{d\theta^2}\log p(\theta|\boldsymbol{x})\right]$ segue que,

$$\log p(\theta|\boldsymbol{x}) \approx \text{constante} \times \exp\left\{-\frac{h(\theta^*)}{2}(\theta-\theta^*)^2\right\}.$$

Portanto, temos a seguinte aproximação para a distribuição a posteriori de θ,

$$\theta|\boldsymbol{x} \sim N(\theta^*, h(\theta^*)^{-1}).$$

Exemplo 4.12 : Seja o modelo,

$$\begin{aligned}X_1,\ldots,X_n &\sim Poisson(\theta) \\ \theta &\sim Gama(\alpha,\beta).\end{aligned}$$

Já vimos que,

$$\theta|\boldsymbol{x} \sim Gama(\alpha+\sum x_i, \beta+n)$$

portanto,

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

ou equivalentemente,

$$\log p(\theta|\boldsymbol{x}) = (\alpha+\sum x_i - 1)\log\theta - \theta(\beta+n) + \text{constante}.$$

A primeira e segunda derivadas são dadas por,

$$\begin{aligned}\frac{d}{d\theta}\log p(\theta|\boldsymbol{x}) &= -(\beta+n) + \frac{\alpha+\sum x_i - 1}{\theta} \\ \frac{d^2}{d\theta^2}\log p(\theta|\boldsymbol{x}) &= -\frac{\alpha+\sum x_i - 1}{\theta^2}.\end{aligned}$$

Segue então que,

$$\theta^* = \frac{\alpha+\sum x_i - 1}{\beta+n}, \quad h(\theta) = \frac{\alpha+\sum x_i - 1}{\theta^2} \quad \text{e} \quad h(\theta^*) = \frac{(\beta+n)^2}{\alpha+\sum x_i - 1}.$$

e a distribuição a posteriori aproximada é,

$$\theta|\boldsymbol{x} \sim N\left(\frac{\alpha+\sum x_i-1}{\beta+n}, \frac{\alpha+\sum x_i-1}{(\beta+n)^2}\right).$$

Para 20 observações simuladas da distribuição de Poisson com parâmetro $\theta = 2$ e hiperparâmetros $a = 1$ e $b = 2$ obteve-se $\sum x_i = 35$. Portanto, a distribuição a posteriori aproximada é,

$$\theta|\boldsymbol{x} \sim N(1.59, 0.07).$$

Na Figura 4.12 estão as densidades a posteriori Gama e usando a aproximação normal.

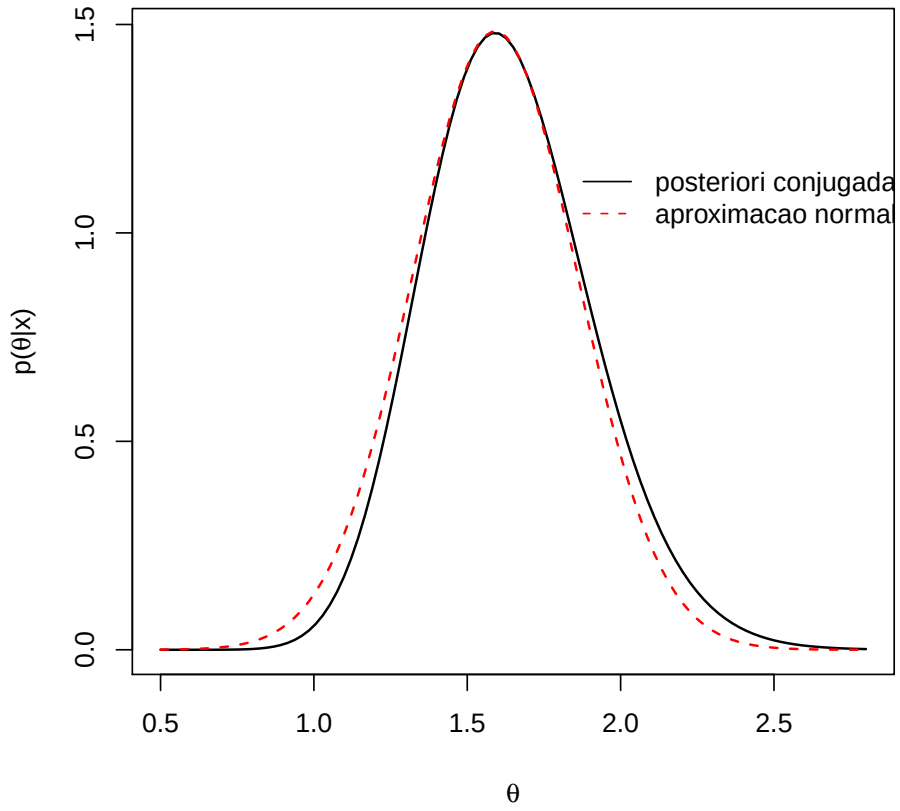

Figura 4.12: Densidades a posteriori conjugada e usando a aproximação normal.