4
Conceitos Básicos de Estatística Bayesiana e Simulação Estocástica

4.1
Elementos de inferência Bayesiana

Gamerman [34] define que tanto o modelo Bayesiano quanto o freqüentista trabalham na presença de observações $\boldsymbol{x}$, cujo valor é inicialmente incerto e descrito através de uma distribuição de probabilidades $f(x|\theta)$. A quantidade θ[4] serve como indexador da família de distribuições das observações representando características de interesse que se deseja conhecer para poder ter uma descrição completa do processo. Continuando, Gamerman [34] alerta para o fato de que o principal interesse de estudo é a determinação do seu valor, não sendo portanto, um simples indexador.

Vale citar que para fazer inferência sobre θ dado x, é necessário determinar uma distribuição de probabilidade conjunta:

$$p(x,\theta) = f(x|\theta)p(\theta) \tag{4.1.1}$$

Após observar $X = x$, pode-se utilizar o Teorema de Bayes para determinar a distribuição a posteriori:

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} \tag{4.1.2}$$

ou

$$p(\theta|x) \propto f(x|\theta)p(\theta) \tag{4.1.3}$$

A razão por inserir a proporcionalidade em (4.1.2) resultando (4.1.3), é que o denominador (constante normalizadora) de (4.1.2) não é função de θ.

O objetivo dos tópicos seguintes é justificar estes preceitos de forma mais detalhada.

[4] Em vários casos θ pode ser multi-dimensional.

4.2
Distribuição a priori

Medrano & Migon [62] reportam que a implementação da inferência segundo a escola Bayesiana requer que as distribuições a priori sejam especificadas para as quantidades de interesse. Neste contexto, Koop [54] sugere que o especialista colha diversas opiniões de profissionais do setor em estudo para, assim, tirar conclusões mais convincentes. No entanto, quando as prioris dos cientistas são fortes e divergentes, Gamerman & Migon [35] citam que é fundamental produzir uma análise neutra visando gerar um referencial.

Vale notar que em algumas situações, o pesquisador pode ter o sentimento de que a informação disponível para avaliar a distribuição a priori não existe. Este conhecimento pobre ou ignorância é tema de muita discussão na literatura. Afinal, é importante ressaltar que nunca se pode estar em completo estado de ignorância. Para expressar a idéia de pouco conhecimento a respeito do parâmetro, uma classe de prioris denominadas não-informativas tem se mostrado bastante útil. De acordo com Gamerman & Migon [35], inicialmente foram propostas prioris distribuídas uniformemente ($p(\theta) \propto k$, $-\infty < \theta < \infty$), o que implica não favorecer nenhum valor particular de θ. Neste caso, qualquer análise Bayesiana é baseada fundamentalmente na distribuição amostral. Entretanto, é importante sublinhar algumas dificuldades com relação a esta escolha:

- $p(\theta)$ é uma distribuição imprópria, ou seja: $\int p(\theta) d\theta \to \infty$, se o intervalo de θ for ilimitado;
- $p(\theta)$ não é invariante à reparametrização.

Um método alternativo para determinar prioris não-informativas, para um conjunto de parâmetros, é conhecido como Regra de Jefrey. Esta regra é baseada na escolha de uma priori proporcional à raiz quadrada do determinante de uma matriz de informação esperada de Fisher. Ehlers [29] explica que o conceito de informação, aqui, está associada a uma espécie de curvatura média da função de verossimilhança no sentido de que quanto maior esta curvatura, mais precisa é a informação contida na verossimilhança.

É digno registrar que esta classe de prioris não-informativas é invariante, porém, eventualmente imprópria. Como, na prática, o objetivo é analisar a distribuição a posteriori, esta constatação é dispensável. No entanto, é de suma importância, antes de realizar qualquer inferência sobre a distribuição a posteriori, certificar que a mesma é uma distribuição própria. Mais uma vez, de acordo com Gamerman & Migon [35], um outro obstáculo dessa técnica é não satizfazer o Princípio da Verossimilhança (este princípio postula que toda informação contida em um determinado experimento, está contida na função de verossimilhança).

Vale notar que muitas vezes pode-se dividir a especificação de uma priori através de estágios. Ehlers [29] explica que, além de facilitar esta especificação, esta abordagem é natural em determinadas situações experimentais.

Para ilustrar, seja a distribuição a priori de θ a qual depende dos valores dos hiperparâmetros ϕ. Pode-se, então, escrever a distribuição condicional $p(\theta|\phi)$ ao invés da marginal $p(\theta)$. Além disso, ao invés de fixar valores para os hiperparâmetros, determina-se uma distribuição a priori $p(\phi)$. Neste ponto, completa-se o segundo estágio da hierarquia. Matematicamente, a distribuição marginal a priori de θ pode ser encontrada através da seguinte integral:

$$p(\theta) = \int p(\theta,\phi)d\phi = \int p(\theta|\phi)p(\phi)d\phi \qquad (4.2.1)$$

Gamerman [34] cita que não há limitação quanto ao número de estágios. Contudo, quanto mais alto o estágio, mais complexo fica a especificação das distribuições. Assim sendo, geralmente são especificados dois (2) ou três (3) estágios.

4.3
Distribuição amostral (Função de Verossimilhança)

Gamerman & Migon [35] mencionam que a função de verossimilhança é a função que associa a cada valor θ, o valor $f(x|\theta)$. Portanto, ao fixar um valor para x e variar os valores de θ, pode ser observado a plausibilidade (ou verossimilhança) de cada um dos valores de θ. Vale lembrar que a inferência Bayesiana obedece ao Principio da Verossimilhança. É através desta função que o conhecimento a priori sobre θ é modificado. Neste ponto, é importante lembrar

que em uma análise científica, é apropriado que esta função seja dominante com relação à distribuição a priori. Afinal, é coerente pensar que qualquer informação relevante sobre o parâmetro θ seja obtida através da realização de um experimento e, consequentemente, através dos dados.

4.4
Distribuição a posteriori

Gelman et al [38] apontam que a distribuição a posteriori contém todas as informações probabilísticas do parâmetro θ. Desta forma, a elaboração de um gráfico para investigar as caracteristicas dessa distribuição pode ser considerado um procedimento bastante útil. Concomitante, pode-se resumir a informação (do parâmetro de interesse) contida na distribuição a posteriori através de valores numéricos (estimativas pontuais) tais como: média, mediana, moda. Ehlers [29] enfatiza que é importante reduzir a incerteza dessas estimativas através do intervalo de credibilidade (intervalo de confiança Bayesiano). Este intervalo de credibilidade expressa a probabilidade de θ estar em um intervalo pré-definido, condicional aos dados observados. Vale lembrar que é possível construir inúmeros intervalos. No entanto, aquele que contempla o menor comprimento possível é denominado de Máxima Densidade a Posteriori (MDP). De acordo com Cespedes [17], MDP é o intervalo em que a densidade para todo ponto pertencente ao intervalo é maior do que para todo não pertencente a ele.

4.5
Exemplo

Este tópico tem como finalidade elucidar as definições abordadas nos tópicos 4.2, 4.3 e 4.4, objetivando uma melhor compreensão das mesmas. Para tal, seja o exemplo registrado em Gamerman & Migon [35] onde dois físicos A e B, visando obter uma estimativa mais acurada de uma constante física θ, concordam em fazer um experimento em laboratório. O físico A é bastante experiente e determina a seguinte distribuição a priori para o parâmetro θ:

$$\theta_A \sim N\left(\mu_A, \tau_A^2\right) \tag{4.5.1}$$

Já o físico B tem pouca experiência (pouco conhecimento) e dessa forma opta por utilizar uma priori não- informativa:

$$\theta_B \sim N\left(\mu_B, \tau_B^2\right) \tag{4.5.2}$$

Vale salientar que estes parâmetros indexadores da família de distribuições a priori (μ_A, μ_B, τ_A^2 e τ_B^2) são denominados na literatura como hiperparâmetros para diferenciá-los dos parâmetros de interesse $\left(\theta_A, \theta_B\right)$. Para este estudo, os valores adotados são os seguintes:

$$\mu_A = 900\,;\ \mu_B = 800\,;\ \tau_A^2 = (20)^2\,;\ \tau_B^2 = (80)^2$$

As curvas na Figura 4.1 mostram estas densidades a priori para A e B:

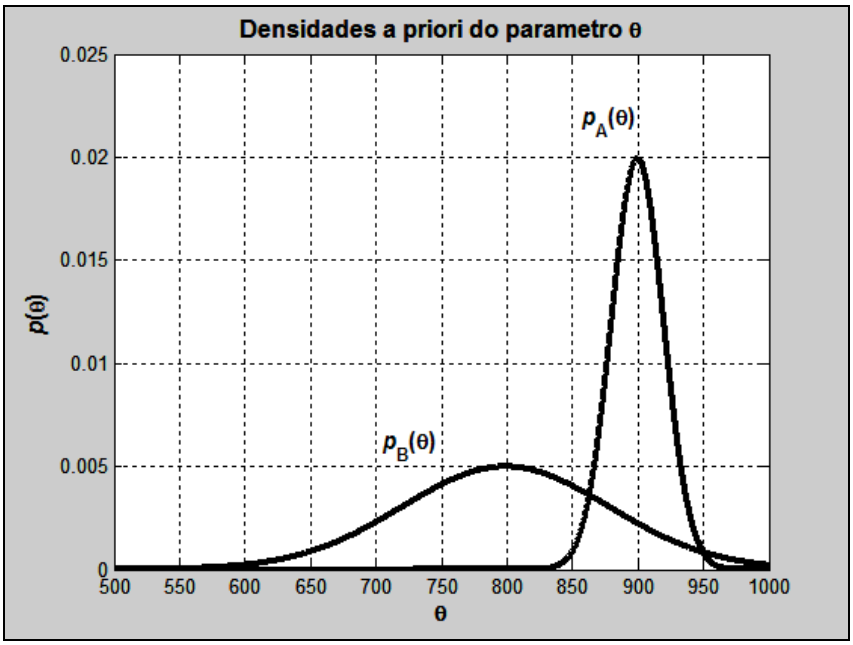

Figura 4.1: Densidades a priori do parâmetro θ.

Sabendo-se que a distribuição amostral $\left(X|\theta\right) \sim N\left(\theta, (40)^2\right)$, tem-se

$$f\left(\boldsymbol{x}\middle|\theta, \sigma^2\right) = l\left(\theta, \sigma^2; \boldsymbol{x}\right) = \prod_{j=1}^{n} f\left(x_j\middle|\theta, \sigma^2\right) \tag{4.5.3}$$

ou

$$l\left(\theta, \sigma^2; \boldsymbol{x}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(x_j - \theta\right)^2\right) \tag{4.5.4}$$

Suponha-se que o resultado de uma simples observação seja $x = 850$; então a função de verossimilhança é mostrada na Figura 4.2:

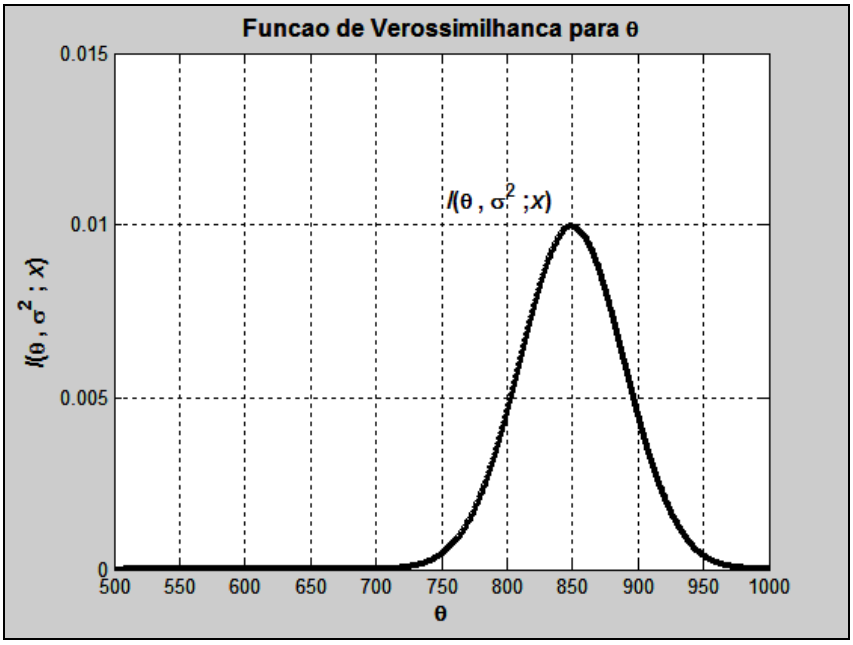

Figura 4.2: Função de Verossimilhança para θ.

A abordagem Bayesiana é adaptativa e permite revisar a distribuição a priori dos parâmetros com novas informações, obtendo-se ao final uma distribuição a posteriori (π). Por inspeção visual da Figura 4.3, pode-se perceber que a aquisição da amostra, e consequentemente a introdução da informação, modifica a distribuição a priori com uma considerável redução com relação à incerteza sobre os parâmetros (θ_A, θ_B).

Dessa forma, tem-se:

- Para o físico A: $(\theta|X = 850) \sim N(890,(17,9)^2)$;
- Para o físico B: $(\theta|X = 850) \sim N(840,(35,7)^2)$.

Estas densidades são mostradas na Figura 4.3:

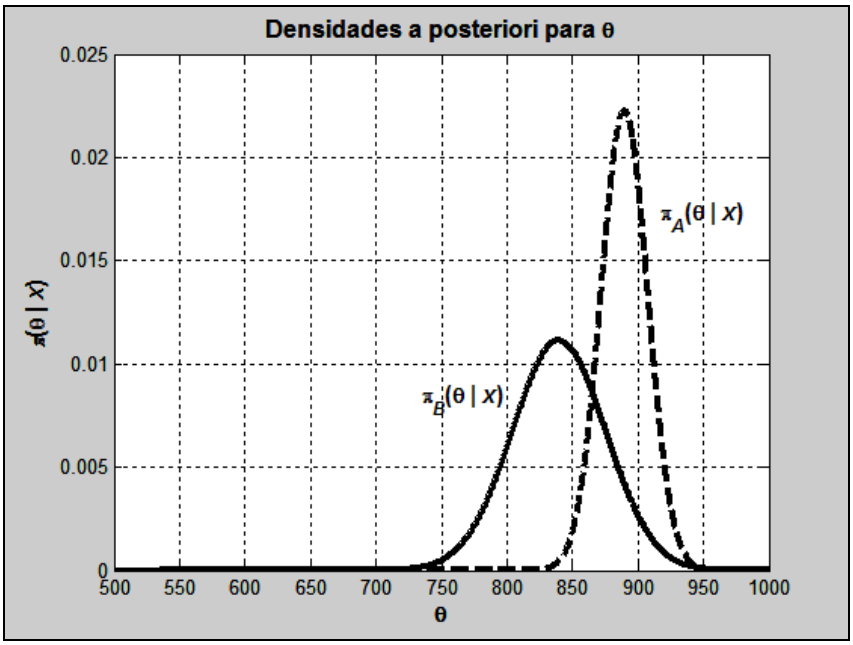

Figura 4.3: Densidades a posteriori do parâmetro θ.

Este exemplo ilustra uma situação importante onde, para um dado modelo observacional, as distribuições a priori e a posteriori pertencem a uma mesma classe de distribuições, no caso a distribuição normal. Esta propriedade de preservação da classe de distribuições define conjugação. Neste caso, a atualização do conhecimento que se tem sobre o parâmetro do modelo envolve apenas uma mudança nos hiperparâmetros.

Gamerman [34] alerta para o cuidado com a utilização indiscriminada de prioris conjugadas. Essencialmente, o problema é que a priori conjugada nem sempre é uma representação correta da incerteza a priori. Sua utilização está, muitas vezes, associada à tratabilidade analítica decorrente.

4.6
Obtenção de resumos de interesse através de simulação

Conforme citado no tópico 4.4, em diversas situações é interessante resumir a informação descrita na distribuição a posteriori em termos de esperanças de funções particulares do parâmetro θ, i.e [29]:

$$\mathbf{E}\left[g(\theta)|X\right] = \int g(\theta)\, p(\theta|x) d\theta \qquad (4.6.1)$$

sendo $g(.)$ uma função qualquer. Por exemplo, se $g(\theta)=\theta$, então $\mathbf{E}[g(\theta)|X]$ representa a média a posteriori.

É digno de registro que estes cálculos de integrais, em geral, não são analiticamente tratáveis. Esta questão, por muitos anos, atrasou o desenvolvimeto de modelos Bayesianos. No entanto, com a evolução computacional, houve uma explosão de trabalhos utilizando métodos baseados em simulação. Os itens seguintes abordam alguns destes métodos: Integração de Monte Carlo, Amostragem por Importância e Monte Carlo via Cadeias de Markov (MCMC). É importante sublinhar que estes textos são de caráter introdutório. Para mais detalhes, o leitor é convidado a consultar as seguintes referências: Gamerman [34], Robert & Casella [76].

4.6.1
Integração via Monte Carlo

De acordo com as informações anteriores, em diversas situações a equação (4.6.1) é altamente complexa. Desta forma, os métodos estatísticos tradicionais falham e as inferências são feitas utilizando simulações. Basicamente, a simulação de Monte Carlo refere-se a qualquer simulação que envolve o uso de números aleatórios.

Em linhas gerais, o objetivo da Integração via Monte Carlo é calcular a seguinte integral:

$$\vartheta = \mathbf{E}(X) = \int x f(x) dx \,. \tag{4.6.1.1}$$

Uma maneira plausível consiste em gerar aleatoriamente amostras $\{X_t, t=1,...,T\}$ da distribuição de X e fazer a aproximação:

$$\mathbf{E}(X) \approx T^{-1} \sum_{t=1}^{T} x_t \tag{4.6.1.2}$$

Este conceito pode ser extrapolado considerando

$$\vartheta = \mathbf{E}(\phi(X)) = \int \phi(x) f(x) dx \,. \tag{4.6.1.3}$$

Novamente, seja $\{X_t, t=1,...,T\}$ uma amostra aleatória independente e identicamente distribuída (*iid*) da distribuição $f(x)$. Então,

$$\hat{\vartheta} = T^{-1} \sum_{t=1}^{T} \phi(x_t) \tag{4.6.1.4}$$

converge para $\mathbf{E}(\phi(X))$ com probabilidade um (1) quando $T \to \infty$. Como $\hat{\vartheta}$ é uma média amostral de observações independentes, temos que $\mathbf{Var}(\hat{\vartheta}) = \frac{\mathbf{Var}(\phi(X))}{T}$. Mas $\mathbf{Var}(\phi(X))$ pode ser estimada através da variância amostral de $\phi(X_t)$, de modo que o erro padrão estimado para $\hat{\vartheta}$ é dado por

$$se(\hat{\vartheta}) = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^{T} (\phi(X_t) - \hat{\vartheta})^2} \ . \tag{4.6.1.5}$$

Para ilustrar, seja o seguinte exemplo apresentado por Roberts [78]. Para este caso, X e Y têm distribuição conjunta (distribuição normal bivariada) dada por:

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_x^2 \sigma_y^2 (1-\rho^2)}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\} \tag{4.6.1.6}$$

sendo ρ o coeficiente de correlação entre X e Y, $-1 < \rho < 1$, $x \in (-\infty, \infty)$, $y \in (-\infty, \infty)$, $(\mu_x, \ \mu_y) \in \mathbf{R}^2$ e $(\sigma_x, \ \sigma_y) \in \mathbf{R}_+^2$.

O valor de $\mathbf{Pr}(X < 1, Y < 1)$ pode ser estimado considerando inicialmente

$$\int I_A(x,y) f(x,y) dxdy \tag{4.6.1.7}$$

e na seqüência calculando:

$$T^{-1} \sum_{t=1}^{T} I_A(x_t, y_t) \tag{4.6.1.8}$$

onde I_A corresponde a uma função indicadora: $A = \{(x,y) : x < 1, y < 1\}$. Há várias maneiras para simular uma distribuição normal bivariada. Conceitualmente, a forma mais fácil é utilizar os métodos baseados em misturas que exploram a questão da densidade conjunta das variáveis aleatórias (X,Y), ou seja:

$$f(x,y) = f(y|x) f(x) \tag{4.6.1.9}$$

Dessa forma, o par (x, y) pode ser gerado em dois (2) passos: inicialmente gera-se $X=x$ a partir da sua distribuição marginal e, então, y é gerado condicionalmente do valor obtido de x, $(Y|X = x)$.

Utilizando $\rho = 0{,}5$, $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$, pode-se obter a seguinte distribuição condicional:

$$f(y|x) = N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y\sqrt{1-\rho^2}\right) \qquad (4.6.1.10)$$

A estimativa da probabilidade requerida após mil (1000) iterações $(T = 1.000)$ é dada por 0,762. Vale registrar que quanto maior for T, mais acurada será a estimativa. A Figura 4.4 fornece o *scatterplot* dos valores simulados:

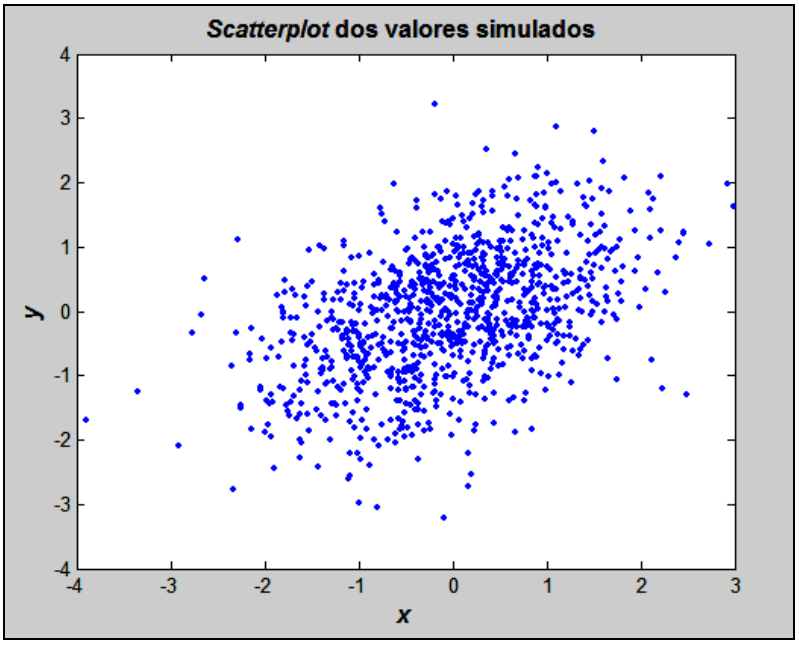

Figura 4.4: *Scatterplot* dos valores simulados.

Pode ser constatado que este método é bastante simples e fácil de se usar mesmo em problemas envolvendo altas dimensões. De acordo com Roberts [78], o custo por esta simplicidade é que a variância é alta.

Uma extensão da Integração via Monte Carlo é dada pela Amostragem por Importância. Esse método é útil quando não é possível gerar amostras de $f(x)$.

Matematicamente, podemos re-escrever a equação (4.6.1.3) como:

$$\vartheta = \mathbf{E}(\phi(X)) = \int \psi(x) g(x) dx \qquad (4.6.1.11)$$

onde $\psi(x) = \frac{\phi(x) f(x)}{g(x)}$ e $g(x)$, denominada de função de importância. A escolha desta função é baseada no fato de que a mesma seja fácil de amostrar. Gamerman [34] mostra que uma escolha ótima, em termos de minimizar a variância é usar $g \propto \phi \text{ x } f$.

Procedendo dessa maneira, o próximo passo consiste em estimar ϑ:

$$\hat{\vartheta} = T^{-1} \sum_{t=1}^{T} \psi(X_t) \tag{4.6.1.12}$$

cuja variância é dada por:

$$\mathbf{Var}(\hat{\vartheta}) = T^{-1} \int (\psi(x) - \vartheta)^2 \, g(x) dx \tag{4.6.1.13}$$

Tierney [91] comenta que, por várias décadas, o método Amostragem por Importância foi utilizado em um contexto Bayesiano para estimar características das distribuições a posteriori tais como a média ou o desvio-padrão de uma função de θ. Entretanto, em diversos casos é impossível amostrar diretamente da distribuição a posteriori ou mesmo descobrir uma função de importância plausível. Desta forma, os métodos baseados em simulação de Monte Carlo tornam-se praticamente inviáveis. Neste cenário, uma poderosa técnica denominada Monte Carlo via Cadeias de Markov (*Markov chain Monte Carlo* - MCMC) tem sido amplamente utilizada para resolver esta espécie de problema.

4.7
Monte Carlo via Cadeias de Markov (MCMC)

Monte Carlo via cadeias de Markov (MCMC) é uma poderosa ferramenta para gerar variáveis aleatórias.

De acordo com Casella & George [16], MCMC foi, inicialmente, proposto por Metropolis et al [65]. Este trabalho pioneiro resultou em desenvolvimentos futuros: Hastings [48], Geman & Geman [40] e Tanner & Wong [86]. Segundo Gilks et al [43], a apresentação da aplicabilidade desses métodos em uma gama de problemas estatísticos convencionais, especificamente em modelagem Bayesiana, é creditado aos artigos de Gelfand & Smith [36] e Gelfand et al [37].

Durante várias décadas, diversos artigos e livros têm sido publicados com a intenção de expor a aplicação dessa metodologia em diversos problemas reais

complexos, como exemplo: Spall [80], Doucet et al [25], Chib & Greenberg [20], Tierney [91], Besag et al [14], Casella & George [16], Gamerman [34], Robert & Casella [76], Gilks et al [41].

O protótipo do problema é como se segue. Considere uma seqüência de variáveis aleatórias $\{X_t, t = 1,...,T\}$ tal que em cada tempo $t \geq 0$, o proximo estado X_{t+1} é gerado condicionalmente da distribuição $P(X_{t+1}|X_t)$, que é denominada *kernel* de transição da cadeia e não depende de t se a cadeia for homogênea no tempo. Isto é, a distribuição de X_{t+1} depende somente do estado atual da cadeia, X_t. Uma realização desta sequência de variáveis aleatórias $\{X_t, t = 1,...,T\}$ é conhecida na literatura como Cadeias de Markov. É importante relatar que X_0 representa alguma condição inicial.

Roberts [77] faz uma breve revisão para que a distribuição de X_t convirja para a distribuição estacionária. Assim sendo, ele explica que a cadeia precisa satisfazer três (3) importantes propriedades:

i. irredutível;
ii. aperiódica;
iii. positiva recorrente.

A propriedade (iii) pode ser considerada a mais importante. Afinal, esta propriedade está relacionada com a existência de uma distribuição estacionária tal que se X_t é gerado desta distribuição, então todos os valores subsequentes também o serão.

Para ilustrar, reporte-se à equação (4.6.1.3) e considere $f(.)$ como sendo a distribuição estacionária. Gilks et al [43] advogam que com o crescimento de t, os pontos amostrados X_t se parecerão mais e mais com amostras dependentes de $f(.)$. Ignorando as primeiras l iterações da cadeia (período denominado "*burn-in*"), pode-se usar a cadeia de Markov para estimar $\mathbf{E}(\phi(X))$, onde $X \sim f(.)$, da seguinte maneira:

$$\bar{\phi} = \frac{1}{T-l} \sum_{t=l+1}^{T} \phi(X_t) \qquad (4.7.1)$$

Tal procedimento é chamado de média ergódica. É importante explicitar que a sua convergência, garantida pelo famoso teorema ergódico (vide Roberts [77]), é quase certa ou com probabilidade 1.

Neste ponto, é necessário descobrir como construir uma cadeia de Markov tal que a sua distribuição estacionária seja precisamente a distribuição de interesse[5]. Os tópicos seguintes têm por finalidade apresentar brevemente dois algoritmos MCMC mais utilizados para este fim: Algoritmo de Metropolis-Hastings e o Amostrador de Gibbs.

4.7.1
Algoritmo de Metropolis-Hastings

De acordo com Haykin, [49], o algoritmo de Metropolis-Hastings, introduzido nos primórdios da ciência da computação e que tem sido utilizado até nos dias atuais pela comunidade Física, baseia-se em um método de Monte Carlo para a simulação estocástica de uma coleção de átomos em equilíbrio a uma dada temperatura.

O algoritmo Metropolis-Hastings é um mecanismo que produz uma cadeia de Markov $\{X_t, t = 1,...,T\}$ com as seguintes características:

- a distribuição limite de interesse (π) pode ser conhecida apenas parcialmente. Assim, a constante normalizadora não precisa ser conhecida, pois será cancelada no quociente;
- nenhuma fatoração de (π) é necessária;
- os métodos são facilmente implementados, e
- a sequência de amostras é obtida através de uma cadeia de Markov.

A partir destas informações, seja a cadeia de Markov $\{X_t, t \geq 0\}$. Para o algoritmo de Metropolis-Hastings, a cada tempo $t \geq 0$ o próximo estado, X_{t+1}, é escolhido, primeiramente, amostrando um ponto candidato Y da distribuição proposta $q(.|X_t)$.

[5] É importante esclarecer que, em muitos textos, a distribuição de interesse (distribuição conjunta a posteriori) é denotada por π.

O ponto candidato Y é, então, aceito como o próximo estado da cadeia com probabilidade dada por:

$$\alpha(X_t, Y) = \mathbf{min}\left\{1, \frac{\pi(Y)q(X_t|Y)}{\pi(X_t)q(Y|X_t)}\right\} \qquad (4.7.1.1)$$

Vale salientar que o uso de (4.7.1.1) é fundamental para garantir que se construa uma cadeia ergódica e que, portanto, se obtenha a distribuição estacionária. Então, se o ponto candidato Y for aceito, o próximo estado será $X_{t+1} = Y$. Caso contrário, ou seja, se o candidato for rejeitado, a cadeia não se moverá, isto é, $X_{t+1} = X_t$.

Vale explicar que quando a distribuição proposta for simétrica ($q(Y|X) = q(X|Y)$), tem-se um caso especial do algoritmo Metropolis-Hastings nomeado de algoritmo Metropolis. Neste caso, a equação (4.7.1.1) é sensivelmente modificada transformando-se em:

$$\alpha(X_t, Y) = \mathbf{min}\left\{1, \frac{\pi(Y)}{\pi(X_t)}\right\} \qquad (4.7.1.2)$$

Em termos práticos, os seguintes passos podem ser utilizados para gerar uma seqüência de amostras aleatórias através do algoritmo Metropolis-Hastings [41]:

i. inicie a cadeia em X_t e ajuste $t = 0$;

ii. gere um ponto candidato $Y \sim q(.|X_t)$;

iii. gere $U \sim Unif(0,1)$;

iv. se $U \leq \alpha(X_t, Y)$ então ajuste $X_{t+1} = Y$, senão ajuste $X_{t+1} = X_t$;

v. incremente $t = t+1$ e repita os passos (ii) até (v).

Para esclarecer este algoritmo, suponha-se que a distribuição de Cauchy padrão, sem a constante normalizadora, seja distribuição limite de interesse:

$$\pi(x) \propto \frac{1}{(1+x^2)}; \qquad -\infty < x < \infty \qquad (4.7.1.3)$$

A distribuição proposta, *q*, para este exemplo é uma normal com a média dada pelo valor prévio e o desvio-padrão atribuído arbitrariamente como $\sigma = 2$.

Os resultados obtidos pela simulação são apresentados nas Figuras 4.5 e 4.6. Vale informar que a Figura 4.5 fornece um gráfico construído com o valor da

variável e o número de iterações. Pode ser verificado que este gráfico está de acordo com Melo & Ehlers [64] que afirmam que o mesmo deve apresentar um caráter aleatório.

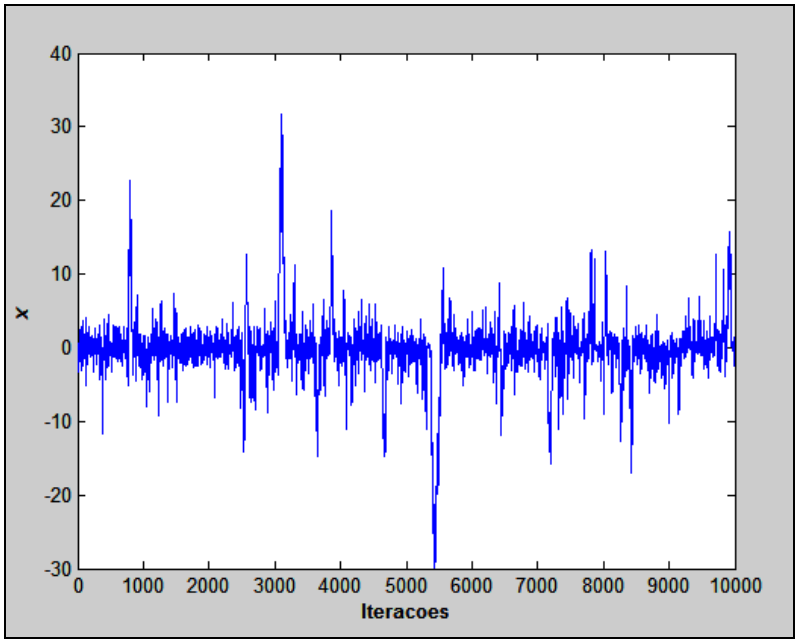

Figura 4.5: Valores simulados através do algoritmo Metropolis-Hastings.

A seguir, a Figura 4.6 mostra o histograma da densidade estimada (aproximada) com a curva correspondente à função densidade de probabilidade verdadeira (equação 4.7.1.3).

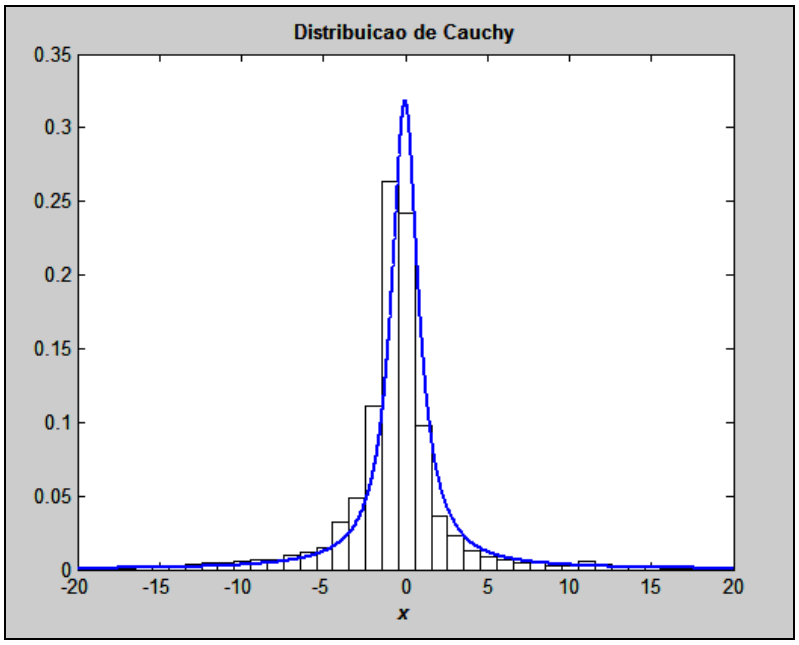

Figura 4.6: Algoritmo Metropolis-Hastings na geração da distribuição de Cauchy.

Não é difícil perceber que, neste exemplo, as amostras poderiam ter sido geradas diretamente da distribuição de Cauchy, sem a necessidade de aplicação do algoritmo Metropolis-Hastings.

4.7.2
Amostrador de Gibbs

O Amostrador de Gibbs foi originalmente concebido dentro de um contexto de reconstrução de imagens por Geman & Geman [40] e está contido em uma grande classe de esquemas de simulação estocástica que utilizam cadeias de Markov [34].

Embora ele seja um caso especial do algoritmo Metropolis-Hastings, o mesmo apresenta duas particularidades, a saber:

- Todos os pontos gerados são aceitos;
- Existe a necessidade de conhecermos a distribuição condicional completa.

Entenda por distribuição condicional completa, a distribuição da *d*-ésima componente de um determinado parâmetro condicionada a todas as outras componentes.

Segundo Gamerman [34], o amostrador de Gibbs é, essencialmente, um esquema iterativo de amostragem de uma cadeia de Markov, cujo *kernel* de transição é formado pelas distribuições condicionais completas.

Para descrever o funcionamento deste algoritmo, suponha que o objetivo seja determinar algumas características da distribuição marginal de uma variável aleatória θ_1 em um problema constituído de mais duas variáveis aleatórias: θ_2 e θ_3. Este tratamento é abordado em Casella & George [16]. Matematicamente, tem-se:

$$p(\theta_1) = \iint p(\theta_1, \theta_2, \theta_3) d\,\theta_2\, d\,\theta_3 \tag{4.7.2.1}$$

Analisando a equação (4.7.2.1), percebe-se que para obter a distribuição marginal, é necessário o cálculo de uma integral sobre todas as outras variáveis. Extrapolando e considerando que em muitas aplicações θ seja um vetor de parâmetros particionado como $\theta = (\theta_1, \ldots, \theta_d)$, a integral (4.7.2.1) torna muito complicada (e algumas vezes impossível) de se avaliar. O amostrador de Gibbs é uma forma de estimar densidades marginais através de simulação. Uma questão que é importante frisar refere-se ao fato de que a partição escolhida seja fácil de amostrar. Neste ponto, é interessante apresentar os procedimentos deste algoritmo [34]:

i. inicie o contador de iterações da cadeia $t = 1$ e arbitre valores iniciais $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$;

ii. obtenha um novo valor de $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})$ a partir de $\theta^{(t-1)}$ através de sucessivas gerações de valores:

$$\theta_1^{(t)} \sim p\left(\theta_1 \middle| \theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}\right)$$
$$\theta_2^{(t)} \sim p\left(\theta_2 \middle| \theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)}\right)$$
$$\vdots$$
$$\theta_d^{(t)} \sim p\left(\theta_d \middle| \theta_1^{(t)}, \ldots, \theta_{d-1}^{(t)}\right)$$

iii. mude o contador de t para $t+1$ e retorne a (ii) até a convergência.

Assim, cada iteração se completa após d movimentos ao longo dos eixos coordenados das componentes de θ. Após a convergência, os valores resultantes

formam uma amostra de $p(\theta)$[6]. Vale notar que, mesmo em problemas envolvendo grandes dimensões, são utilizadas simulações univariadas ou em bloco o que, em geral, é uma vantagem computacional [29]. Este fato tem contribuído de forma significativa para a aplicação desta metodologia, principalmente na área de econometria aplicada com ênfase Bayesiana (vide, por exemplo, Koop et al [57]).

Para ilustrar, seja um processo de Poisson com priori hierárquica proposto por Ehlers [30]:

$$Y_j \sim \text{Poi}(\lambda),\ j = 1,...,n \tag{4.7.2.2}$$

$$\lambda \sim \text{Exp}(\beta) \tag{4.7.2.3}$$

$$\beta \sim \text{Gama}(\text{C},\text{D}) \tag{4.7.2.4}$$

A partir dessas informações, pode-se escrever a equação da distribuição conjunta de Y, λ e β como:

$$p(\boldsymbol{y},\lambda,\beta) = \prod_{j=1}^{n} p(y_j|\lambda)p(\lambda|\beta)p(\beta) \tag{4.7.2.5}$$

Ao observar $Y_j = y_j$ e fazendo $z = \sum_{j=1}^{n} y_j$, tem-se as seguintes distribuições condicionais completas:

$$\pi(\lambda|\boldsymbol{y},\beta) \propto \mathbf{exp}[-\lambda(\beta+n)]\lambda^z \tag{4.7.2.6}$$

$$\pi(\beta|\boldsymbol{y},\lambda) \propto \mathbf{exp}[-\beta(\lambda+\text{D})]\beta^{\text{C}} \tag{4.7.2.7}$$

Então, as distribuições condicionais completas são dadas por:

$$\pi(\lambda|\boldsymbol{y},\beta) \sim \text{Gama}(z+1,\beta+n) \tag{4.7.2.8}$$

$$\pi(\beta|\boldsymbol{y},\lambda) \sim \text{Gama}(1+\text{C},\lambda+\text{D}) \tag{4.7.2.9}$$

Em seguida, é realizada uma simulação com cem (100) dados $(n = 100)$ de um processo com média aproximadamente quatro (4), $\text{C} = \text{D} = 0{,}1$. Por fim, rodou-se mil (1000) iterações $(T = 1.000)$. A Figura 4.7 fornece os resultados da simulação:

[6] a distribuição *p* não precisa, necessariamente, ser uma distribuição a posteriori [16].

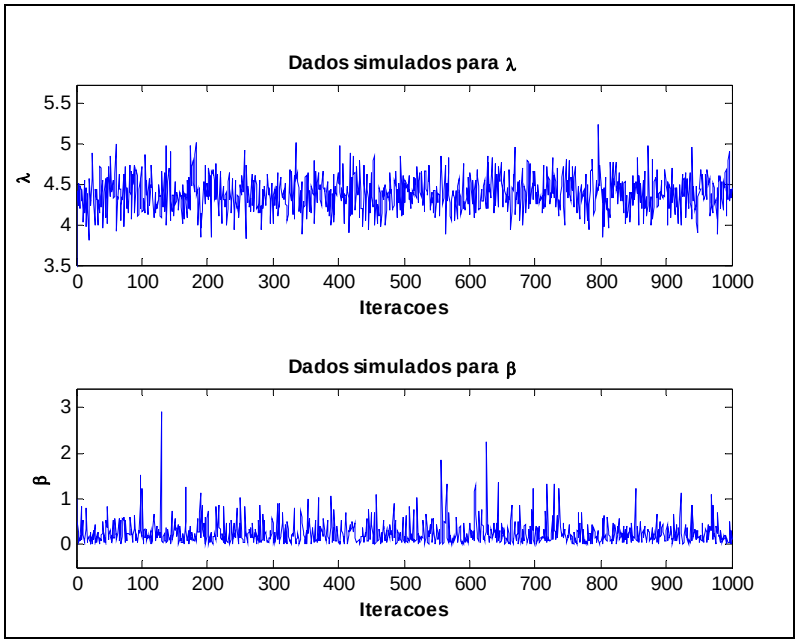

Figura 4.7: Dados simulados para λ e β utilizando o algoritmo de Gibbs.

Descartando-se as duzentas (200) primeiras iterações $(l = 200)$, é possível calcular as médias amostrais do parâmetro λ ao longo das iterações. A Figura 4.8 representa graficamente os resultados obtidos. Por inspeção visual da mesma, verifica-se que é notória a convergência do mesmo.

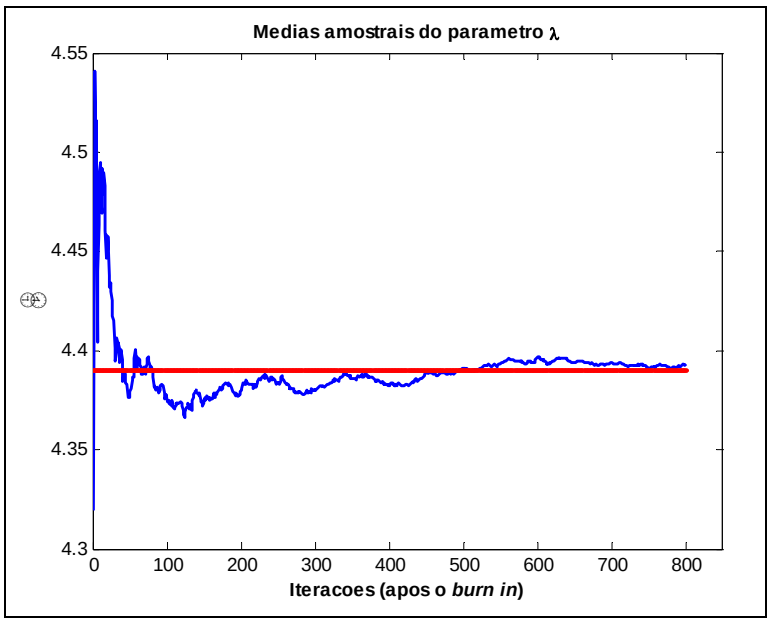

Figura 4.8: Resultado das médias amostrais para o parâmetro λ.

Agora, suponha que o objetivo seja estimar a distribuição marginal da variável aleatória X. Para tal, seja a seguinte distribuição conjunta de X, Y e N [16]:

$$p(x,y,n) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}\,\mathbf{exp}(-\lambda)\frac{\lambda^n}{n!} \qquad (4.7.2.10)$$

$$x = 0,1,...,n\,;\ 0 \le y \le 1\,;\ n = x, x+1,...$$

Como a distribuição marginal de X não tem forma fechada, faz-se necessário o uso de simulação para a determinação da mesma. Utilizando o mesmo raciocínio do primeiro exemplo, têm-se as seguintes distribuições condicionais completas:

$$\pi(x|y,n) \sim \text{Bin}(n,y) \qquad (4.7.2.11)$$

$$\pi(y|x,n) \sim \text{Beta}(x+\alpha, n-x+\beta) \qquad (4.7.2.12)$$

$$\pi(n|x,y) \propto \mathbf{exp}[-\lambda(1-y)]\frac{[\lambda(1-y)]^{n-x}}{(n-x)!} \qquad (4.7.2.13)$$

A Figura 4.9 mostra a distribuição estimada considerando que os valores utilizados foram: $\lambda = 16$, $\alpha = 2$, $\beta = 4$, $T = 5.000$ e $l = 200$.

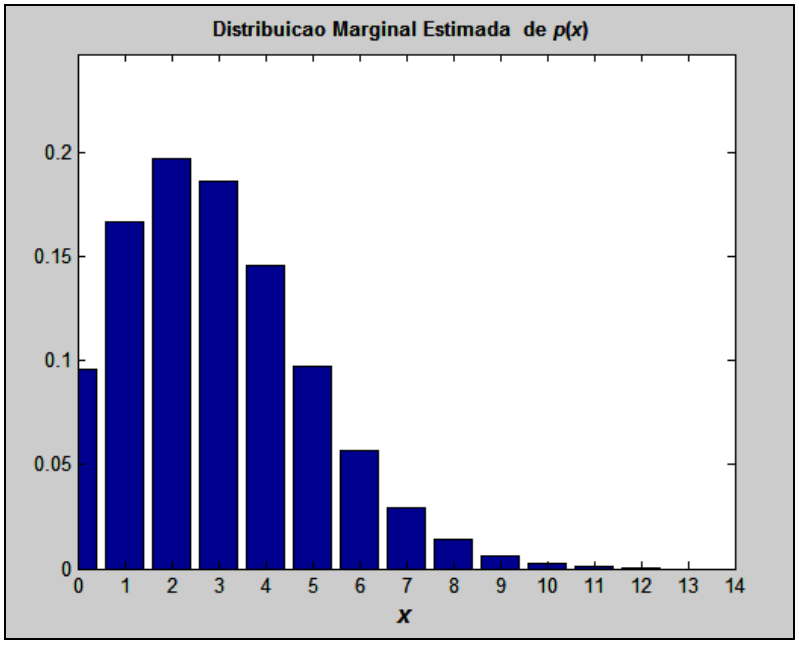

Figura 4.9: Distribuição marginal estimada de $p(x)$.

Até o momento, pôde ser constatado que a implementação dos algoritmos MCMC é relativamente simples. Esta facilidade constantemente tem resultado em erros crassos, dentre os quais Gelman [38] cita:

1. modelagem inapropriada;
2. erros de programação;
3. baixa convergência dos parâmetros do modelo.

Embora todos os três (3) itens citados sejam de suma importância, o terceiro tópico é um assunto que tem despertado muito interesse, principalmente na comunidade estatística com enfoque Bayesiano, onde esses algoritmos têm sido exaustivamente utilizados.

4.7.3
Monitoração da convergência

Uma vez que a simulação é realizada, é necessário verificar a convergência das seqüências simuladas. O excelente trabalho de Cowles & Carlin [22] faz uma revisão e também compara diversos métodos propostos na literatura para este fim. Ainda de acordo com Cowles & Carlin [22], os mesmos podem ser concentrados em duas (2) grandes áreas:

i. Teórica: esta área envolve o conhecimento de uma matemática sofisticada como também uma infinidade de cálculos laboriosos que precisam ser repetidos para cada modelo sob certas considerações;
ii. Aplicada: esta área foca nos resultados produzidos pela simulação.

Um método bastante simples, para monitorar a convergência, é o gráfico de autocorrelação amostral. Neste gráfico é interessante que o decaimento das barras seja o mais rápido possível. No entanto, muitas das vezes, devido à natureza Markoviana das amostras colhidas, isso não acontece. Nestes casos, algumas soluções propostas são [36,42, 74]:

a) Reparametrização;
b) *Thinning*;
c) Cadeias Paralelas;

d) Aumentar o número de iterações da cadeia e, consequentemente, o *burn-in.*

A Figura 4.10, gerada a partir das simulações realizadas no primeiro exemplo do tópico 4.7.2, mostra os gráficos de autocorrelação amostral para os parâmetros λ e β.

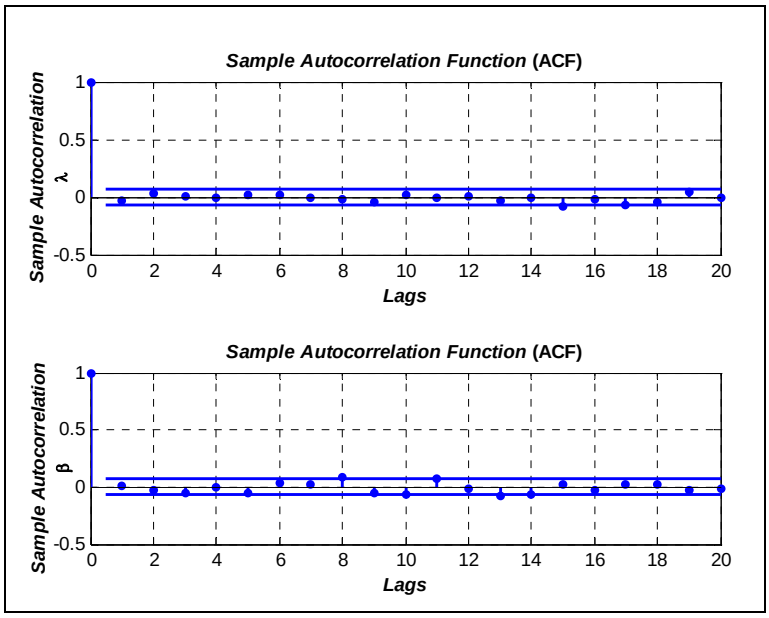

Figura 4.10: Gráficos de autocorrelação amostral dos parâmetros λ e β.

Pela característica exposta, ou seja, decaimento rápido, não é difícil conjecturar a convergência dos parâmetros e, neste caso, que as suas estimativas são confiáveis.