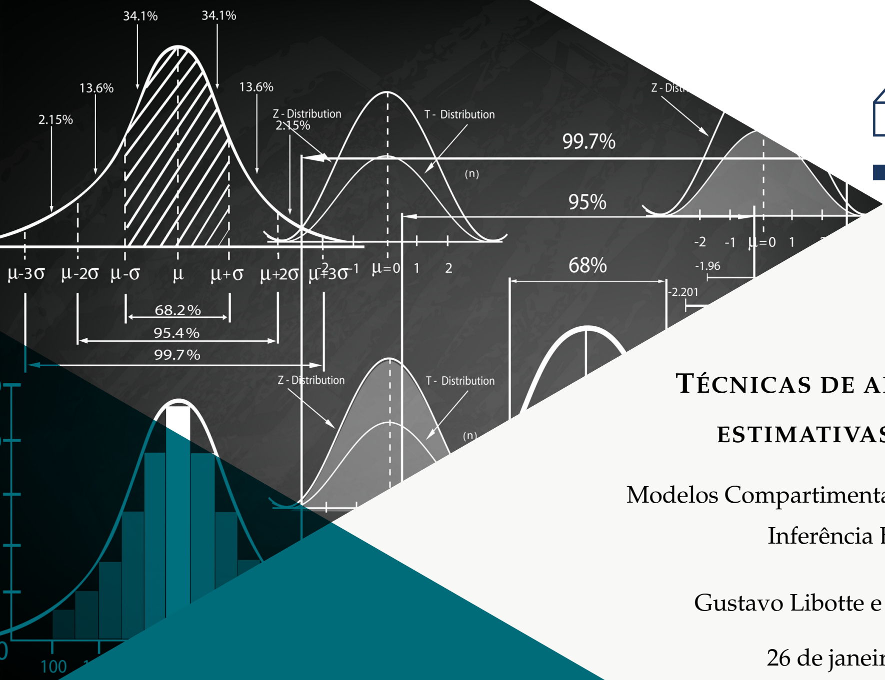


Laboratório Nacional de Computação Científica
UNIDADE DE PESQUISA DO MCTI

Técnicas de amostragem e estimativas pontuais

Modelos Compartimentais em Epidemiologia e Inferência Bayesiana

Gustavo Libotte e Regina Almeida

26 de janeiro de 2022

O Algoritmo de Metropolis
Relembrando

- Tendo obtido um valor de proposta θ^*, nós o adicionamos ou uma cópia de $\theta^{(s)}$ ao nosso conjunto, dependendo da proporção $r = p\left(\theta^* \mid y\right) / p\left(\theta^{(s)} \mid y\right)$.
- Especificamente, dado $\theta^{(s)}$, o algoritmo de Metropolis gera um valor $\theta^{(s+1)}$ como segue:
    1. Gere uma amostra $\theta^* \sim J\left(\theta \mid \theta^{(s)}\right)$.
    2. Calcule a taxa de aceitação:
    $$r = \frac{p\left(\theta^* \mid y\right)}{p\left(\theta^{(s)} \mid y\right)} = \frac{p\left(y \mid \theta^*\right) p\left(\theta^*\right)}{p\left(y \mid \theta^{(s)}\right) p\left(\theta^{(s)}\right)}$$
    3. Tome
    $$\theta^{(s+1)} = \begin{cases} \theta^*, \text{ com probabilidade } \min(r, 1) \\ \theta^{(s)}, \text{ com probabilidade } 1 - \min(r, 1) \end{cases}$$
- A etapa 3 pode ser realizada amostrando $u \sim \text{Uniforme}(0, 1)$ e definindo $\theta^{(s+1)} = \theta^*$ se $u < r$, ou definindo $\theta^{(s+1)} = \theta^{(s)}$ caso contrário.

O Algoritmo de Metropolis
Um exemplo mais detalhado

- Vamos experimentar o algoritmo de Metropolis para o modelo normal conjugado com uma variância conhecida, uma situação em que **sabemos** a distribuição *a posteriori* correta.
- Tomando $\theta \sim \text{Normal}\left(\mu, \tau^2\right)$ e $\{y_1, \ldots, y_n \mid \theta\} \overset{i.i.d.}{\sim} \text{Normal}\left(\theta, \sigma^2\right)$, a distribuição *a posteriori* de θ é Normal $\left(\mu_n, \tau_n^2\right)$ onde

$$\mu_n = \bar{y}\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} + \mu\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}$$
$$\tau_n^2 = 1/\left(n/\sigma^2 + 1/\tau^2\right).$$

Derivação de μ_n e τ_n^2

Você pode ver a derivação do modelo normal conjugado no Apêndice A.3 de [1, p. 241].

[1] S. K. Ghosh and B. J. Reich.
Bayesian statistical methods.
Chapman & Hall/CRC, Boca Raton, 1 edition, 2019.

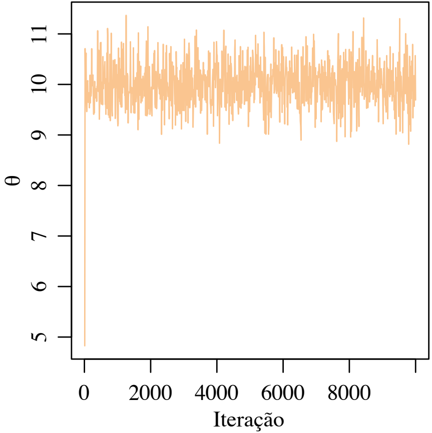
O Algoritmo de Metropolis
Um exemplo mais detalhado

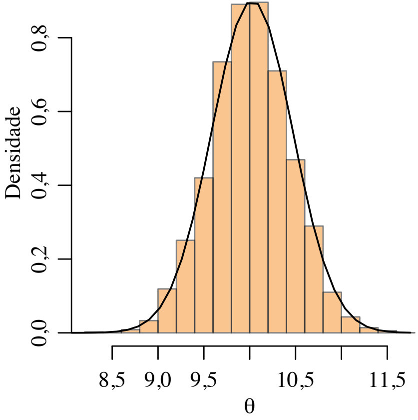
- Suponha que $\sigma^2 = 1, \tau^2 = 10, \mu = 5, n = 5$ e $\boldsymbol{y} = (9{,}37,\ 10{,}18,\ 9{,}16,\ 11{,}60,\ 10{,}33)$.
- Para estes dados, $\mu_n = 10{,}03$ e $\tau_n^2 = 0{,}20$.
- Com base neste modelo e na distribuição *a priori*, a taxa de aceitação comparando um valor proposto θ^* a um valor atual $\theta^{(s)}$ é

$$r = \frac{p\left(\theta^* \mid \boldsymbol{y}\right)}{p\left(\theta^{(s)} \mid \boldsymbol{y}\right)}$$

- Em muitos casos, calcular a relação r diretamente pode ser **numericamente instável**, um problema que muitas vezes pode ser remediado calculando o logaritmo de r.
- Mantendo as coisas na escala logarítmica, a proposta é aceita se $\log u < \log r$, onde u é uma amostra da distribuição uniforme em $(0, 1)$.
- No resultado a seguir, foram geradas 10.000 iterações do algoritmo de Metropolis, começando em $\theta^{(0)} = 0$ e usando uma distribuição de proposta normal, $\theta^{(s+1)} \sim \text{Normal}\left(\theta^{(s)}, \delta^2\right)$ com $\delta^2 = 2$.

O Algoritmo de Metropolis
Um exemplo mais detalhado





O Algoritmo de Metropolis
Um exemplo mais detalhado

- O algoritmo de Metropolis gera uma sequência dependente $\left\{\theta^{(1)}, \theta^{(2)}, \ldots\right\}$ de θ-valores.
- Como nosso procedimento para gerar $\theta^{(s+1)}$ depende apenas de $\theta^{(s)}$, a distribuição condicional de $\theta^{(s+1)}$ dado $\left\{\theta^{(1)}, \ldots, \theta^{(s)}\right\}$ também depende apenas de $\theta^{(s)}$ e portanto da sequência $\left\{\theta^{(1)}, \theta^{(2)}, \ldots\right\}$ é uma **cadeia de Markov**.
- Para qualquer valor numérico dado θ_a de θ,
$$\lim_{S \to \infty} \frac{\#\left\{\theta\text{'s na sequência} < \theta_a\right\}}{S} = p\left(\theta < \theta_a \mid y\right).$$
- Isso sugere que podemos aproximar médias, quantis e outras quantidades *a posteriori* de interesse usando a distribuição empírica de $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\}$.
- No entanto, nossa aproximação a essas quantidades dependerá de quão bem nossa sequência simulada realmente se aproxima de $p(\theta \mid y)$.
- Os resultados da teoria da probabilidade dizem que, no limite de $S \to \infty$, a aproximação será exata, mas **na prática** não podemos executar a cadeia de Markov para sempre.

O Algoritmo de Metropolis
Um exemplo mais detalhado

- Em vez disso, a prática padrão na aproximação MCMC, usando o algoritmo de Metropolis (ou o amostrador de Gibbs), é a seguinte:
  1. execute o algoritmo até alguma iteração B para a qual parece que a cadeia de Markov atingiu a **estacionariedade**;
  2. execute o algoritmo S mais vezes, gerando $\left\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\right\}$;
  3. descarte $\left\{\theta^{(1)}, \ldots, \theta^{(B)}\right\}$ e use a distribuição empírica de $\left\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\right\}$ para aproximar $p(\theta \mid y)$.
- As iterações até e incluindo B são chamadas de período de "*burn-in*", no qual a cadeia de Markov se move de seu valor inicial para uma região do espaço de parâmetros que tem **alta probabilidade *a posteriori***.
- Se tivermos uma boa ideia de onde está essa região de alta probabilidade, podemos reduzir o período de "*burn-in*" iniciando a cadeia de Markov daquele ponto.
- Neste exemplo, teria sido melhor começar com $\theta^{(1)} = \bar{y}$.
- No entanto, começar com $\theta^{(1)} = 0$ ilustra que o algoritmo é capaz de se mover de uma região de baixa probabilidade *a posteriori* para uma de alta probabilidade.

Estimativa de máxima verossimilhança

- Considere o problema de estimar um conjunto de parâmetros θ de um modelo probabilístico, dado um conjunto de observações $x_1, x_2, \ldots, x_n$.
- As técnicas de máxima verossimilhança assumem que
  1. As amostras não dependem umas das outras, em que a ocorrência de um não tem efeito sobre os outros.
  2. Cada um deles pode ser modelado exatamente da mesma maneira.
- Isso significa que os eventos são independentes e identicamente distribuídos (i.i.d.).
- A suposição sobre i.i.d. implica que um modelo para a função densidade de probabilidade conjunta para todas as observações consiste no produto do mesmo modelo de probabilidade $p\left(x_i \mid \theta\right)$ aplicado a cada observação independentemente.
- Para n observações, isso pode ser escrito como

$$p\left(x_1, x_2, \ldots, x_n \mid \theta\right) = p\left(x_1 \mid \theta\right) p\left(x_2 \mid \theta\right) \ldots p\left(x_n \mid \theta\right)$$

- Cada função $p\left(x_i \mid \theta\right)$ tem os mesmos valores de parâmetro θ, e o objetivo da estimativa de parâmetro é maximizar um modelo de probabilidade conjunta desta forma.

Estimativa de máxima verossimilhança

- Como as observações não mudam, este valor só pode ser alterado alterando a escolha dos parâmetros θ.

$$L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Como os dados são fixos, é sem dúvida mais útil pensar nisso como uma função de verossimilhança para os parâmetros, que somos livres para escolher.
- Multiplicar muitas probabilidades pode levar a números muito pequenos e, portanto, as pessoas geralmente trabalham com o logaritmo da probabilidade, ou log-verossimilhança:

$$\ln L(\theta \mid x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln p(x_i \mid \theta),$$

- Como os logaritmos são funções estritamente crescentes monotonicamente, maximizar a probabilidade logarítmica é o **mesmo** que maximizar a verossimilhança:

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i \mid \theta).$$

Máximo *a posteriori*

- A máxima verossimilhança assume que todos os valores de parâmetros são **igualmente prováveis** *a priori*: não julgamos alguns valores de parâmetros como mais prováveis do que outros antes de considerarmos as observações.
- Em vez de simplesmente computar a estimativa de máxima verossimilhança, ainda podemos obter alguns dos benefícios da abordagem bayesiana ao permitir que a distribuição *a priori* **influencie** a escolha da estimativa pontual.
- Uma maneira racional de fazer isso é escolher a estimativa do ponto **máximo *a posteriori*** (MAP).
- A estimativa MAP escolhe o ponto de probabilidade *a posteriori* máxima (ou densidade de probabilidade máxima no caso mais comum de θ contínuo):

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta \mid x) = \arg\max_{\theta} \{\ln p(x \mid \theta) + \ln p(\theta)\}$$

- Observe que, para uma *priori* **uniforme**, o termo $\ln p(\theta)$ é uma constante e a expressão acima coincide então com a solução de máxima verossimilhança.
- Discussão: **Como computar $\theta_{\mathbf{ML}}$ e $\theta_{\mathbf{MAP}}$ de forma eficiente?**