MAT02034 - Métodos bayesianos para análise de dados

Introdução a computação bayesiana

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Métodos Computacionais

Métodos Computacionais

- Como visto, a inferência bayesiana é baseada na aplicação do teorema de Bayes

$$g(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)g(\theta)}{\int_{\Theta} f(\boldsymbol{x}|\theta)g(\theta)d\theta} = c(\boldsymbol{x})f(\boldsymbol{x}|\theta)g(\theta) \propto f(\boldsymbol{x}|\theta)g(\theta),$$

e na obtenção de **medidas resumo** dessa distribuição, como $E[\theta|\boldsymbol{x}]$, intervalos de credibilidade ou probabilidades *a posteriori*.

Métodos Computacionais

- A maior dificuldade na aplicação de inferência bayesiana está justamente no cálculo das integrais envolvidas, tanto no cálculo de $f(\boldsymbol{x})$ para a obtenção da distribuição *a posteriori*, quanto na obtenção das medidas resumos citadas anteriormente.
- Devido a isso, a inferência bayesiana ganhou muito força com o avanço computacional das últimas décadas.
- Nesta aula discutimos alguns recursos que podem ser utilizados na inferência bayesiana.

Métodos Computacionais

- Muitos dos métodos descritos baseiam-se na *Lei dos Grande Números* (LGN).

Lei dos Grande Números

Seja $X_1, X_2, \ldots$ uma sequência de variáveis aleatórias $i.i.d.$ com $\mathrm{E}\left[X_1\right] = \mu$ e $\mathrm{Var}\left[X_1\right] = \sigma^2 < \infty$, então

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu.$$

Métodos Computacionais

- As integrais de interesse serão escritas como o valor esperado de funções de variáveis aleatórias

$$\int h(x)dF(x) = \mathrm{E}\left[h(X)\right].$$

- Deste modo, suponha que $X_1, X_2, \ldots$ é uma sequência de variáveis aleatórias $i.i.d.$ e $h : \mathbb{R} \longrightarrow \mathbb{R}$ é uma função tal que $Var\left[h(X_1)\right] < \infty$. Então, pela *LGN*,

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \longrightarrow \mathrm{E}\left[h(X_1)\right].$$

Método de Monte Carlo

Método de Monte Carlo

- Suponha que deseja-se calcular

$$\int_{\Theta} h(\theta)g(\theta|\boldsymbol{x})d\theta = \mathrm{E}_g\left[h(\theta)|\boldsymbol{x}\right]$$

e é possível **simular** realizações $(i.i.d.)$ $\{\theta_1, \ldots, \theta_M\}$ da distribuição *a posteriori* $g(\theta|\boldsymbol{x})$.

- Então, a integral acima pode ser aproximada por

$$\frac{1}{M}\sum_{i=1}^{M} h(\theta_i).$$

Método de Monte Carlo

- A **precisão da aproximação** é usualmente estimada pelo **erro padrão** da estimativa

$$EP\left[\frac{1}{M}\sum_{i=1}^{M} h(\theta_i)\right] \approx \sqrt{\frac{1}{M}\left(\frac{1}{M}\sum_{i=1}^{M}[h(\theta_i)]^2 - \left[\frac{1}{M}\sum_{i=1}^{M} h(\theta_i)\right]^2\right)}.$$

Método de Monte Carlo: aproximando π

- Suponha que deseja-se estimar o número π usando o método de Monte Carlo. Considere então que o vetor aleatório (X, Y) tem distribuição conjunta uniforme em um quadrado centrado na origem, $\mathfrak{X} = [-1, 1] \times [-1, 1]$, e um círculo A de raio 1 inscrito nesse quadrado, $x^2 + y^2 \leq 1$.

Método de Monte Carlo: aproximando π

Método de Monte Carlo: aproximando π

- Como a distribuição é uniforme no quadrado, a probabilidade de escolher um ponto no círculo é

$$\Pr(A) = \frac{\text{área da círculo}}{\text{área do quadrado}} = \frac{\pi}{4} = \int_A f(x,y)dxdy = \int_{\mathfrak{X}} \mathbb{I}_A(x,y)\frac{1}{4}\,dxdy = \mathrm{E}\left[\mathbb{I}_A(X,Y)\right].$$

- Suponha que é possível gerar uma amostra $\{(x_1,y_1),\ldots,(x_M,y_M)\}$ de (X,Y), de modo que podemos aproximar o valor de π por

$$\pi = 4\Pr(A) = \mathrm{E}\left[4\mathbb{I}_A(X,Y)\right] \approx \frac{1}{M}\sum_{i=1}^{M} 4\mathbb{I}_A(x_i,y_i),\ x_i \overset{iid}{\sim} Unif(-1,1),\ y_i \overset{iid}{\sim} Unif(-1,1).$$

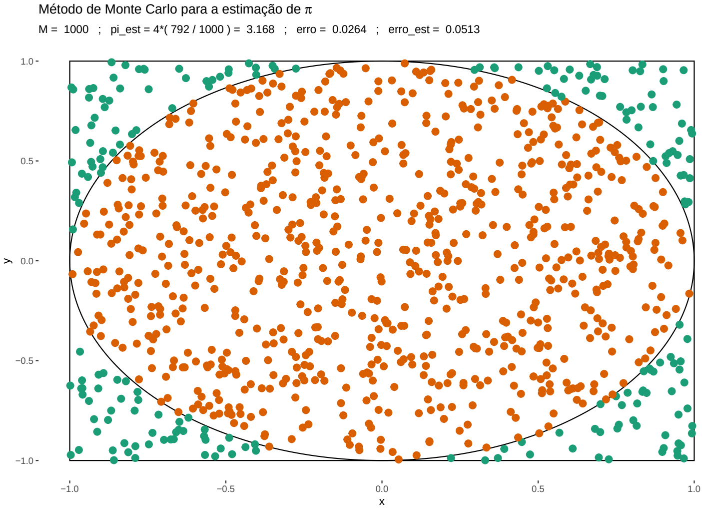
- No R: `x <- runif(M, -1, 1); y <- runif(M, -1, 1).`

Método de Monte Carlo: aproximando π

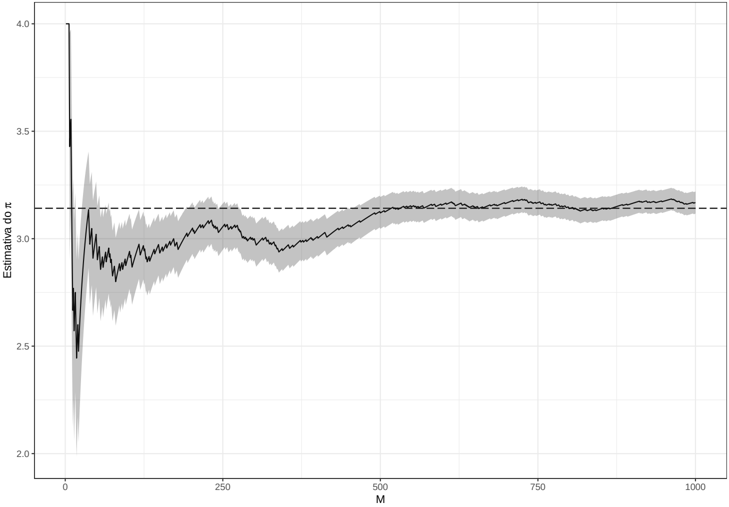
- Denotando por $t = \sum_{i=1}^{M} \mathbb{I}_A(x_i, y_i)$, o erro estimado é

$$
\begin{aligned}
&\sqrt{\frac{1}{M}\left(\frac{1}{M}\sum_{i=1}^{M}\left[4\mathbb{I}(x_i, y_i)\right]^2 - \left[\frac{1}{M}\sum_{i=1}^{M}4\mathbb{I}(x_i, y_i)\right]^2\right)} \\
&= \sqrt{\frac{1}{M}\left(\frac{16}{M}t - \left[\frac{4}{M}t\right]^2\right)} \\
&= \sqrt{\frac{16}{M}\frac{t}{M}\left(1 - \frac{t}{M}\right)} \\
&\leq \sqrt{\frac{16}{M}\frac{1}{4}} = \frac{2}{\sqrt{M}}.
\end{aligned}
$$

Método de Monte Carlo: aproximando π



Método de Monte Carlo: aproximando π



Método de Monte Carlo: cálculo

- Suponha que não sabemos que

$$\int_0^1 x^3(1-x)^5 e^x dx = 74046 - 27240e \approx 0.0029928.$$

- O método de Monte Carlo pode ser utilizado para estimar (aproximar) esta integral.

Método de Monte Carlo: cálculo

Vamos considerar duas estratégias:

1. $U \sim Unif(0,1)$ e a integral pode ser escrita como $\mathrm{E}\left[U^3(1-U)^5 e^U\right]$.

Ou seja,

$$\mathrm{E}\left[U^3(1-U)^5 e^U\right] \approx \frac{1}{M}\sum_{i=1}^{M}[u_i^3(1-u_i)^5 e^{u_i}],\ u_i \overset{iid}{\sim} Unif(0,1).$$

- No R: `u <- runif(M, 0, 1)`.

Método de Monte Carlo: cálculo

2. $Y \sim Beta(4,6)$ de modo que

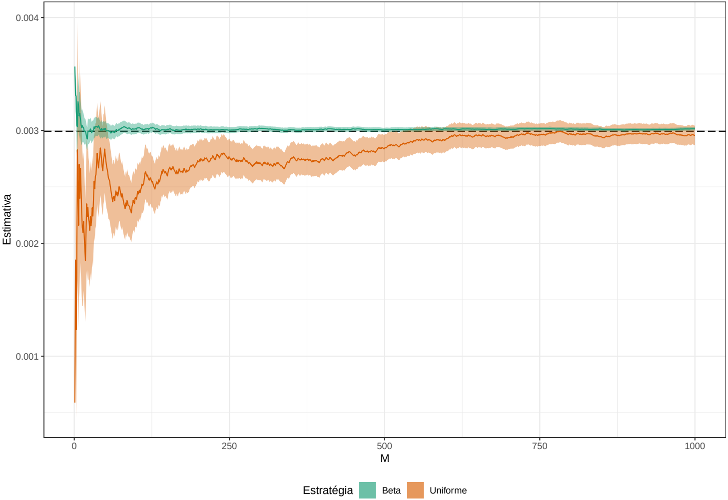
$$\int_0^1 y^3(1-y)^5 e^y dy = B(4,6)\int_0^1 e^y \frac{y^{4-1}(1-y)^{6-1}}{B(4,6)}\, dy = B(4,6)E\left[e^Y\right].$$

Ou seja,

$$B(4,6)E\left[e^Y\right] \approx B(4,6)\times\frac{1}{M}\sum_{i=1}^{M}[e^{y_i}],\ \ y_i \overset{iid}{\sim} Beta(4,6)$$

- No `R`: `y <- rbeta(M, 4, 6)`.

Método de Monte Carlo: cálculo



Método de Monte Carlo: aplicações em inf. bayesiana

- Em 1786 Laplace estava interessado em determinar se a probabilidade θ de um nascimento masculino em Paris durante um certo período de tempo era superior a 0.5 ou não.
- Os números oficiais forneceram $y_1 = 251527$ nascimentos do sexo masculino para $y_2 = 241945$ nascimentos do sexo feminino.
    - A proporção observada foi, portanto, de 0,509.

Método de Monte Carlo: aplicações em inf. bayesiana

- Escolhemos uma distribuição uniforme como distribuição *a priori* para θ a proporção de nascimentos do sexo masculino.
  - A distribuição *a posteriori* é

$$g(\theta|y) = Beta(\theta; 251528, 241946).$$

Método de Monte Carlo: aplicações em inf. bayesiana

- Imagine que não temos a tabela (da distribuição Beta) e estamos interessados na **média *a posteriori*** dessa distribuição.
- Além disso, imagine que podemos amostrar (usando um computador) um grande número M de amostras independentes $(\theta_i, i = 1, \ldots, M)$ dessa distribuição.

Método de Monte Carlo: aplicações em inf. bayesiana

- Pode-se propor o seguinte estimador

$$\frac{1}{M}\sum_{i=1}^{M}\theta_i$$

como pela lei dos grandes números,

$$\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M}\theta_i = \mathrm{E}_{g(\theta|y)}(\theta).$$

Método de Monte Carlo: aplicações em inf. bayesiana

- Também podemos estimar a **variância *a posteriori***, pois

$$\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} \theta_i^2 = \mathrm{E}_{g(\theta|y)}(\theta^2).$$

Método de Monte Carlo: aplicações em inf. bayesiana

- Agora considere os seguintes problemas mais desafiadores:
  - queremos encontrar estimativas da mediana dessa distribuição *a posteriori*, bem como um intervalo de credibilidade de 95%.
- Começamos com a mediana e assumimos que ordenamos as amostras, ou seja, para qualquer $i < j, \theta_i < \theta_j$ e por simplicidade que M é um número par.

Método de Monte Carlo: aplicações em inf. bayesiana

- Seja $\overline{\theta}$ a mediana da distribuição *a posteriori*. Então sabemos que

$$
\begin{aligned}
\Pr(\theta_i \geq \overline{\theta}) &= \int_{-\infty}^{\infty} \mathbb{I}(\overline{\theta} < \theta) g(\theta|y) d\theta = 1/2 \\
\Pr(\theta_i \leq \overline{\theta}) &= \int_{-\infty}^{\infty} \mathbb{I}(\overline{\theta} > \theta) g(\theta|y) d\theta = 1/2
\end{aligned}
$$

de modo que (assumindo por simplicidade que M é par e que ordenamos $(\theta_i, i = 1, \ldots, M)$), é sensato escolher uma estimativa para $\overline{\theta}$ entre $\theta_{N/2}$ e $\theta_{N/2+1}$.

Método de Monte Carlo: aplicações em inf. bayesiana

- Agora suponha que estamos procurando por θ^- e θ^+ tais que

$$\Pr(\theta^- \leq \theta \leq \theta^+) = \int_{-\infty}^{\infty} \mathbb{I}(\theta^- \leq \theta \leq \theta^+) g(\theta|y) d\theta = 0.95$$

ou

$$\Pr(0 \leq \theta \leq \theta^-) = 0.025 \quad \text{e} \quad \Pr(\theta^+ \leq \theta \leq 1) = 0.025$$

e assumindo novamente por simplicidade que $M = 1000$ e que as amostras foram ordenadas. Descobrimos que uma estimativa razoável de θ^- está entre θ_{25} e θ_{26} e uma estimativa de θ^+ entre θ_{975} e θ_{976}.

Método de Monte Carlo: aplicações em inf. bayesiana

- Por fim, podemos estar interessados em calcular

$$
\begin{aligned}
\Pr(\theta < 0.5) &= \int_0^{0.5} g(\theta|y)d\theta \\
&= \int_0^{1} I(\theta \leq 0.5)g(\theta|y)d\theta = \mathrm{E}_{g(\theta|y)}[I(\theta \leq 0.5)] \approx \frac{1}{M}\sum_{i=1}^{M} I(\theta_i \leq 0.5).
\end{aligned}
$$

Método de Monte Carlo: considerações

- Método baseado no poder computacional.
- Se sabemos gerar da **distribuição alvo**, facilmente obtemos estimativas de interesse através de resumos amostrais (médias, proporções, estatísticas de ordem, etc.).
  - Aplicação de resultados probabilísticos: *LGN*, *TCL*.
- A Alternativa: integração numérica determinística
  - Veja o `help` das funções em R `area` e `integrate`.
  - Funciona bem em dimensões baixas (unidimensional).
  - Geralmente precisa de algum conhecimento da função.
- **Problema** o que fazer quando não for possível gerar diretamente da distribuição alvo?

Para casa

- Rodar os códigos dos exemplos de aula.
  - Trazer as dúvidas para o Fórum Geral do Moodle e para a próxima aula.
- Exercício (Moodle).

Próxima aula

- Introdução a computação bayesiana (continuação):
  - Amostragem por importância;
  - Método da rejeição;
  - SIR.

Por hoje é só!

Bons estudos!