ESTATÍSTICA COMPUTACIONAL

Ralph dos Santos Silva

http://www.ralphsilva.org/estatisticacomputacional.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Ementa

- Solução de sistemas lineares.
- Métodos de decomposição de matrizes.
- Problemas de otimização e maximização de funções.
- Geração de variáveis aleatórias.
- Integração por Monte Carlo.
- Métodos de quadratura.
- Aproximação de Laplace.
- Simulação estocástica via cadeias de Markov.

Referências

- Doucet, A., de Freitas, N. e Gordon, N. (2001). Sequential Monte Carlo in Practice. Springer.
- Gamerman, D. e Lopes, H. F. (2006). Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference. 2nd Ed, Chapman & Hall.
- Golub, G.H. e van Loan, C.F. (2012). Matrix Computations. 4th Ed., Johns Hopkins University Press.
- Liu, J. S. (2004). Monte Carlo Strategies in Scientific Computing. Springer.
- Robert, C.P. and Casella, G. (2004). Monte Carlo Statistical Methods. 2nd Ed., Springer.
- Gentle, J.E., Härdle, W.K. e Mori, Yuichi (2012). Handbook of Computational Statistics - Concepts and Methods.

Álgebra linear

Muitos métodos estatísticos exigem a solução de equações, manipulação de matrizes, etc.

Por exemplo, o método de mínimos quadrados se reduz a solução de um sistema de equações lineares.

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} && \text{equação da regressão,} \\ \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b} &= \boldsymbol{X}^\top \boldsymbol{y} && \text{equações normais,} \\ \boldsymbol{b} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} && \text{solução de mínimos quadrados.} \end{aligned}$$

O método de componentes principais é baseado em autovalores $\boldsymbol{\Delta} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ e autovetores $\boldsymbol{P}$ (colunas), $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top$.

Para calcular a função de densidade de probabilidade da normal multivariada precisamos calcular o determinante e inverter a matriz de covariâncias $\boldsymbol{\Sigma}$ (uma matriz positiva definida).

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right),$$

com $\boldsymbol{x} \in \mathbb{R}^d$ e $\boldsymbol{\mu} \in \mathbb{R}^d$.

Matriz positiva definida

Em muitas aplicações importantes usamos matrizes positivas definidas.

Da **Análise de Regressão**, se $\boldsymbol{X}$ tem posto completo p, então $\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{X}$ é positiva definida. Ainda temos que $\det(\boldsymbol{A}) \neq 0$ e que $\boldsymbol{A}$ possui inversa.

Uma matriz $\boldsymbol{A}$ é dita positiva definida se $\boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{c} > 0$ para qualquer vetor $\boldsymbol{c}$ diferente de $\mathbf{0}$.

Uma matriz $\boldsymbol{A}$ é dita positiva semi-definida se $\boldsymbol{c}^\top \boldsymbol{A}\boldsymbol{c} \geqslant 0$ para qualquer vetor $\boldsymbol{c}$ diferente de $\mathbf{0}$.

Matrizes de covariâncias (por exemplo, da distribuição normal multivariada) são sempre simétricas e positivas definidas.

Uma matriz $\boldsymbol{A}$ é dita ser simétrica se $\boldsymbol{A} = \boldsymbol{A}^\top$.

Solução de sistemas lineares

O problema de resolver o sistema linear $\boldsymbol{Ax} = \boldsymbol{b}$ é central em computação.

$\boldsymbol{A}$ é a matriz de coeficientes, e $\boldsymbol{x}$ representa a solução de interesse.

Se $n < m$ (número de equações menor que o número de incognitas), ou se $n \geqslant m$, mas $\boldsymbol{A}$ não tem posto completo (algumas equações são combinações de outras) o sistema é dito indeterminado.

Se $n > m$ e $\boldsymbol{A}$ tem posto completo, então o sistema é superdeterminado e não há solução.

Se $m = n$ e $\boldsymbol{A}$ tem posto completo, então a solução é única.

Abordaremos sistemas em que $m = n$ e $\boldsymbol{A}$ tem posto completo.

Duas abordagens:

1. métodos diretos que obtém solução exata, exceto pelos erros numéricos; e
2. métodos iterativos que obtém solução aproximada.

Sistemas triangulares

Em geral, para resolver um sistema devemos transformá-lo em triangular, isto é, $\boldsymbol{Lx} = \boldsymbol{b}$.

Substituição para frente:

$$\begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Se $\ell_{11}\ell_{22} \neq 0$, então as quantidades desconhecidas x_1 e x_2 são:

$$\ell_{11}x_1 = b_1 \Rightarrow x_1 = \frac{b_1}{\ell_{11}} \quad \text{e} \quad \ell_{21}x_1 + \ell_{22}x_2 = b_2 \Rightarrow x_2 = \frac{b_2 - \ell_{21}x_1}{\ell_{22}}. \qquad (1)$$

Logo, x_i é obtido sequencialmente. Para uma matriz 3×3, temos que

$$\begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

Para $\ell_{11}\ell_{22}\ell_{33} \neq 0$, segue-se que x_1 e x_2 são dados pela Equação (1) e x_3 por

$$\ell_{31}x_1 + \ell_{32}x_2 + \ell_{33}x_3 = b_3 \Rightarrow x_3 = \frac{b_3 - \ell_{32}x_2 - \ell_{31}x_1}{\ell_{33}}.$$

Generalizando para dimensão $n \geqslant 2$, temos

$$\begin{aligned} x_1 &= \frac{b_1}{\ell_{11}} \\ x_i &= \frac{b_i - \sum_{j=1}^{i-1} \ell_{ij}x_j}{\ell_{ii}}, \quad i = 2, \ldots, n. \end{aligned}$$

Ver arquivo exemplo_01.r

No caso de substituição para trás, temos o mesmo algoritmo para uma matriz triangular superior.

Isto é, resolvemos $\boldsymbol{Ux} = \boldsymbol{b}$, em que $\boldsymbol{U}$ é uma matriz triangular superior e tem zeros abaixo da diagonal.

Note que os algoritmos de solução de sistemas em geral usam matrizes triangulares. Mais tarde veremos alguns métodos para decompor matrizes.

Métodos iterativos

Assuma que $\boldsymbol{A}$ é uma matriz $n \times n$ e que o sistema de interesse, $\boldsymbol{Ax} = \boldsymbol{b}$, é bem determinado.

Métodos iterativos mais comuns: **Jacobi, Gauss-Seidel, relaxamento sucessivo, gradiente, etc**.

Seja $\boldsymbol{D}$, $\boldsymbol{L}$ e $\boldsymbol{U}$ as partes diagonal, inferior e superior da matriz $\boldsymbol{A}$ tal que

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}.$$

Ideia do método

Um método iterativo para resolver $\boldsymbol{Ax} = \boldsymbol{b}$ forma uma série de valores $\boldsymbol{x}_i$, para $i = 1, 2, \ldots,$ tal que sob certas condições, essa série converge para a solução exata do sistema.

Um vetor inicial $\boldsymbol{x}_0$ geralmente é escolhido como uma aproximação de $\boldsymbol{x}$ (mas uma má escolha não causa divergência do algoritmo).

Regra para atualizar a série: $\boldsymbol{x}_{i+1} = \boldsymbol{B}_i\boldsymbol{x}_i + \boldsymbol{C}_i\boldsymbol{b}$, com $\boldsymbol{C}_i$ e $\boldsymbol{B}_i$ matrizes $n \times n$.

Diferentes escolhas de $\boldsymbol{B}_i$ e $\boldsymbol{C}_i$ levam a diferentes métodos.

Consideraremos métodos estacionários em que $\boldsymbol{B}_i = \boldsymbol{B}$ e $\boldsymbol{C}_i = \boldsymbol{C}$.

Dessa forma a regra é $\boldsymbol{x}_{i+1} = \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{C}\boldsymbol{b}$ com algumas condições impostas para convergência do método.

Condição: $\boldsymbol{B} + \boldsymbol{CA} = \boldsymbol{I}$.

O método pode ser parado de acordo com critérios pré-especificados. Por exemplo, podemos usar $||\boldsymbol{x}_i - \boldsymbol{x}_{i+1}|| < \epsilon$. Ou podemos usar um número máximo de iterações.

Método de Jacobi

Seja $\boldsymbol{A}$ uma matriz com diagonal diferente de 0 (zero). Então, $\boldsymbol{A}$ pode ser reescrita como

$$\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}.$$

Dessa forma, o sistema $\boldsymbol{Ax} = \boldsymbol{b}$ pode ser reescrito como

$$\boldsymbol{Dx} + (\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x} = \boldsymbol{b}.$$

Isto implica em

$$\boldsymbol{x} = \boldsymbol{D}^{-1}[(-\boldsymbol{L} - \boldsymbol{U})\boldsymbol{x} + \boldsymbol{b}].$$

Temos, então, um método iterativo para encontrar a solução do sistema

$$\boldsymbol{x}_{i+1} = -\boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{U})\boldsymbol{x}_i + \boldsymbol{D}^{-1}\boldsymbol{b}.$$

Finalmente, temos que

$$x_{k,i+1} = \frac{1}{A_{kk}} \left(b_k - \sum_{j=1, j \neq k}^{n} A_{kj} x_{j,i} \right).$$

Ver arquivo exemplo_02.r

Decomposição de Cholesky

Originalmente desenvolvida para resolver problemas de mínimos quadrados em topografia.

Serve para decompor uma matriz positiva definida $\boldsymbol{A}$ em duas matrizes triangulares, tal que $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$.

$\boldsymbol{U}$ é conhecida como Cholesky de $\boldsymbol{A}$ e a relação $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ é conhecida como decomposição de Cholesky.

Teorema

Seja $\boldsymbol{A}$ uma matriz $n \times n$ simétrica e positiva definida. Então, existe uma única matriz triangular superior $\boldsymbol{U}$ com elementos positivos na diagonal tal que $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$.

Decomposição de Cholesky

Algoritmo

Para $i = 1, \ldots, n$

$$U_{ii} = \begin{cases} A_{ii}^{1/2}, & \text{se } i = 1; \\ \left(A_{ii} - \sum_{k=1}^{i-1} U_{ki}^2 \right)^{1/2}, & \text{se } i > 1. \end{cases}$$

Para $j = i + 1, \ldots, n$

$$U_{ij} = \left(A_{ij} - \sum_{k=1}^{i-1} U_{ki} U_{kj} \right) / U_{ii}.$$

Essa decomposição permite a solução do sistema linear $\boldsymbol{Ax} = \boldsymbol{b}$.

No passo 1 resolvemos o sistema $\boldsymbol{U}^\top \boldsymbol{z} = \boldsymbol{b}$ para encontrar $\boldsymbol{z}$.

No passo 2 resolvemos o sistema $\boldsymbol{Ux} = \boldsymbol{z}$ para encontrar $\boldsymbol{x}$.

Ver arquivo exemplo_03.r

Decomposição LU (“lower upper”)

Reduz a matriz $\boldsymbol{A}$ em $\boldsymbol{A} = \boldsymbol{LU}$, em que $\boldsymbol{L}$ é triangular inferior e $\boldsymbol{U}$ é triangular superior.

Ao contrário da decomposição de Cholesky não exige que $\boldsymbol{A}$ seja positiva definida.

Teorema
Seja $\boldsymbol{A}$ uma matriz $n \times n$ que possui decomposição $\boldsymbol{LU}$ tal que $\boldsymbol{A} = \boldsymbol{LU}$ se $\det(\boldsymbol{A}(1:k, 1:k))$ é diferente de 0 para $k = 1, \ldots, n-1$. Se a fatorização $\boldsymbol{LU}$ existe e $\boldsymbol{A}$ é não singular, então a fatorização $\boldsymbol{LU}$ é única e $\det(\boldsymbol{A}) = U_{11} \ldots U_{nn}$.

Decomposição LU (“lower upper”)

Algoritmo

- Faça $\boldsymbol{U} = \boldsymbol{A}$ e $\boldsymbol{L} = \boldsymbol{I}$.
- Para $i = 1, \ldots, (n-1)$ e para $j = i+1, \ldots, n$
  - (a) Faça $L_{ji} = U_{ji} / U_{ii}$.
  - (b) Para $k = i, \ldots, n$, faça $U_{jk} = U_{jk} - L_{ji} U_{ik}$.

Essa decomposição permite a solução do sistema linear $\boldsymbol{Ax} = \boldsymbol{b}$.

No passo 1 resolvemos o sistema $\boldsymbol{Lz} = \boldsymbol{b}$ para encontrar $\boldsymbol{z}$.

No passo 2 resolvemos o sistema $\boldsymbol{Ux} = \boldsymbol{z}$ para encontrar $\boldsymbol{x}$.

Ver arquivo exemplo_04.r

Matriz inversa

A matriz $\boldsymbol{B}$ é dita matriz inversa de $\boldsymbol{A}$ se $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$, em que $\boldsymbol{I}$ é a matriz identidade.

Se $\boldsymbol{B}$ existe, $\boldsymbol{A}$ é dita não singular.

Uma matriz quadrada é singular se, e somente se, $\det(\boldsymbol{A}) = 0$.

Matrizes singulares são raras no sentido de que se escolhemos uma matriz aleatoriamente com distribuição uniforme nas entradas da matriz, então quase certamente a matriz é não singular.

- Consideramos matrizes quadradas ($n \times n$).
- Se a matriz $\boldsymbol{A}$ é de posto completo, então $\det(\boldsymbol{A}) \neq 0$.
- Matriz $\boldsymbol{A}$ simétrica: $\boldsymbol{A} = \boldsymbol{A}^\top$.
- Matriz $\boldsymbol{A}$ idempotente: $\boldsymbol{AA} = \boldsymbol{A}$.

Exemplos podem ser obtidos de **Análise de Regressão**: $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$.

Temos $\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. Segue-se que $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ e $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}$ tal que $\widehat{\boldsymbol{y}} = \boldsymbol{Py}$ e $\boldsymbol{e} = \boldsymbol{My}$. Temos que $\boldsymbol{P}$ e $\boldsymbol{M}$ são matrizes simétricas, idempotentes e ortogonais ($\boldsymbol{PM} = \boldsymbol{MP} = \boldsymbol{0}$).

Considere o problema de inverter uma matriz positiva definida $\boldsymbol{A}$.

Podemos escrever $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ sendo $\boldsymbol{U}$ uma matriz triangular superior que pode ser obtida pela decomposição de Cholesky.

A matriz inversa $\boldsymbol{X}$ é dada por: $\boldsymbol{X}\boldsymbol{U} = \boldsymbol{I}$.

Neste caso, podemos usar o método da substituição para trás para resolver um sistema de equações lineares.

Algoritmo

- Faça $\boldsymbol{X} = \boldsymbol{0}$.
- Para $i = 1, \ldots, n$,
  - (a) $x_{ii} = 1/u_{ii}$.
  - (b) Para $j = i+1, \ldots, n$, $x_{ij} = -\sum_{k=i}^{j-1} \frac{x_{ik} u_{kj}}{u_{jj}}$.

Após obter $\boldsymbol{X}$ (a inversa de $\boldsymbol{U}$), podemos obter a inversa de $\boldsymbol{A}$ usando a propriedade $\boldsymbol{A}^{-1} = \boldsymbol{U}^{-1}(\boldsymbol{U}^\top)^{-1}$.

Ver arquivo exemplo_05.r

Solução de equações não lineares

Considere o problema de resolver uma equação da forma $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$.

Método de Newton-Raphson

Podemos aproximar a função $\boldsymbol{f}(\boldsymbol{x})$ por uma aproximação de Taylor de primeira ordem dada por

$$\boldsymbol{f}(\boldsymbol{x}) \simeq \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{J}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0),$$

sendo $\boldsymbol{J}(\boldsymbol{x}) = \dfrac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}}$ a matriz jacobiano.

No caso de solução de sistemas, teremos $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$ na solução de interesse. Isso leva ao algoritmo iterativo:

$$\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i-1)} - [\boldsymbol{J}(\boldsymbol{x}^{(i-1)})]^{-1} \boldsymbol{f}(\boldsymbol{x}^{(i-1)}).$$

Exemplo

Deseja-se encontrar a raiz quadrada de um número $b \geqslant 0$.

Procuramos a solução $f(x) = 0$ para $f(x) = x^2 - b$.

Sabemos que $f'(x) = 2x$. Daí, temos que

$$x^{(i)} = x^{(i-1)} - \frac{1}{2x^{(i-1)}} \times [(x^{(i-1)})^2 - b].$$

Ver arquivo exemplo_06.r

Exemplo

Agora, deseja-se encontrar $\boldsymbol{x} = (x_1, x_2, x_3)^\top$ que resolve o sistema

$$\begin{cases} f_1(\boldsymbol{x}) &= \exp(x_1) - 2 \\ f_2(\boldsymbol{x}) &= 5x_3 - 4 \\ f_3(\boldsymbol{x}) &= 4x_1x_2 - 2x_3 - 6. \end{cases}$$

tal que

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \exp(x_1) & 0 & 0 \\ 0 & 0 & 5 \\ 4x_2 & 4x_1 & -2 \end{bmatrix}.$$

Ver arquivo exemplo_07.r

Otimização de funções

O problema de encontrar mínimos e máximos de uma função $\boldsymbol{g}(\boldsymbol{x})$ pode ser visto como a solução de um sistema de equações da forma

$$\boldsymbol{u}(\boldsymbol{x}) = \frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{0},$$

em que $\boldsymbol{u}(\boldsymbol{x})$ é o vetor de primeiras derivadas (gradiente). Dessa forma,

$$\boldsymbol{J}(\boldsymbol{x}) = \frac{\partial^2 \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} = \boldsymbol{H}(\boldsymbol{x})$$

é a matriz hessiana (matriz de segundas derivadas). Isso leva ao algoritmo iterativo:

$$\boldsymbol{x}^{(i)} = \boldsymbol{x}^{(i-1)} - [\boldsymbol{H}(\boldsymbol{x}^{(i-1)})]^{-1} \boldsymbol{u}(\boldsymbol{x}^{(i-1)}).$$

Exemplo

Considere o seguinte modelo de regressão Poisson com uma covariável:

$$(y_i|\beta_1, \beta_2) \sim \text{Poi}(\exp(\beta_1 + \beta_2 x_i)), \quad i = 1, \ldots, n.$$

Como encontrar a estimativa de máxima verossimilhança para os coeficientes de interesse?

A função log-verossimilhança é dada por

$$\ell(\beta_1, \beta_2) = c + \sum_{i=1}^{n} \left[y_i(\beta_1 + \beta_2 x_i) - \exp(\beta_1 + \beta_2 x_i) \right].$$

Daí, temos

$$u_1(\beta_1, \beta_2) = \frac{\partial \ell(\beta_1, \beta_2)}{\partial \beta_1} = \sum_{i=1}^{n} [y_i - \exp(\beta_1 + \beta_2 x_i)]$$

$$u_2(\beta_1, \beta_2) = \frac{\partial \ell(\beta_1, \beta_2)}{\partial \beta_2} = \sum_{i=1}^{n} [(y_i - \exp(\beta_1 + \beta_2 x_i))x_i].$$

Qual a solução do sistema $\boldsymbol{u}(\boldsymbol{\beta}) = \mathbf{0}$? Isto é, quais são as estimativas de máxima verossimilhança?

Podemos calcular a matriz de segundas derivadas:

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \begin{bmatrix} \exp(\beta_1 + \beta_2 x_i) & \exp(\beta_1 + \beta_2 x_i) x_i \\ \exp(\beta_1 + \beta_2 x_i) x_i & \exp(\beta_1 + \beta_2 x_i) x_i^2 \end{bmatrix}.$$

Isso leva ao algoritmo iterativo:

$$\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}^{(i-1)} - [\boldsymbol{H}(\boldsymbol{\beta}^{(i-1)})]^{-1} \boldsymbol{u}(\boldsymbol{\beta}^{(i-1)}).$$

Ver arquivo exemplo_08.r

Nota: existe uma variação do algoritmo de Newton-Raphson que utiliza a matriz de informação de Fisher no lugar da matriz hessiana. Procure por este algoritmo (escore de Fisher)!

Simulação

O que é simulação?

Reprodução de um sistema (problema real) em um ambiente controlado pelo pesquisador.

O sistema em questão pode ser determinístico, neste caso, as componentes e os comportamentos do sistema são conhecidos.

Esse sistema pode ser descrito por uma regra matemática exata.

Em outros casos, o sistema está sujeito a flutuações aleatórias.

Então, precisamos acomodar as incertezas envolvidas utilizando a teoria das probabilidades.

Simulação estocástica

No caso de sistemas não determinísticos, existe pelo menos uma componente estocástica, que será descrita por distribuições de probabilidade.

Podemos representar o sistema por

$$y = g(x_1, \ldots, x_p).$$

A função g pode ser complicada.

Os valores $x_1, \ldots, x_p$ são realizações das variáveis aleatórias $X_1, \ldots, X_p$.

Objetivo: reproduzir essas variáveis aleatórias.

Geração de variáveis aleatórias

Essas variáveis aleatórias são, de forma geral, independentemente distribuídas de acordo com uma funcão de distribuição F.

F pode não ser explicitamente conhecida.

Inicialmente, consideraremos geração de variáveis com distribuição uniforme no intervalo (0,1).

Gerando da uniforme

Primeiro, geraremos da $\mathcal{U}(0,1)$ porque essa distribuição fornece a noção básica de representação da aleatoriedade.

E ela será usada para geração de outras distribuições de probabilidade.

Objetivo: gerar uma amostra aleatória $(U_1, \ldots, U_n)$ de tamanho n de $\mathcal{U}(0,1)$.

Paradoxo: queremos gerar números aleatórios, mas usando um mecanismo determinístico que possa ser repetido para gerar vários números aleatórios com a propriedade desejada.

Definição (gerador uniforme pseudo-aleatório): algoritmo que começando de um valor inicial u_0 e uma transformação D, produz uma sequência $\{u_i\} = \{D_i(u_0)\}$ de valores em (0,1). Para todo n, os valores $(u_1, \ldots, u_n)$ reproduzem o comportamento de uma amostra aleatória $(U_1, \ldots, U_n)$ de $\mathcal{U}(0,1)$ quando testadas segundo um certo conjunto de testes estatísticos.

Periódicos: esses algoritmos geram um sequência de números que voltam a se repetir após um período de tamanho S. O algoritmo **Mersenne-Twister**, implementado na grande maioria dos programas, tem um período $S = 2^{19937} - 1$.

Queremos encontrar uma sequência de valores tal que ao testar $H_0 : \{ U_1, \ldots, U_n$ são variáveis aleatórias $\mathcal{U}(0,1)\}$ não rejeitamos H_0.

Teste clássico nesse contexto: Kolmogorov-Smirnov.

Partimos da noção de que queremos encontrar um sistema determinístico que possa imitar um fenômeno aleatório.

Isso sugere que modelos “caóticos” poderiam criar esse comportamento.

Esses modelos são baseados em estruturas complexas tal como sistemas dinâmicos da forma $X_{n+1} = D(X_n)$. Esse tipo de sistema é bastante sensível ao valor inicial.

Exemplo: função logística.

Seja $D_a(x) = ax(1-x)$. Para $a \in [3,57;\ 4,00]$ temos configurações caóticas.

Em particular, para $a = 4,00$ temos uma sequência de números no [0,1] com distribuição arco-seno cuja função de densidade de probabilidade é $\frac{1}{\pi\sqrt{x(1-x)}}$.

Podemos transformar X gerado da distribuição arco-seno e obter amostras da $\mathcal{U}(0,1)$ fazendo: $U_i = \frac{2}{\pi}\text{arcoseno}(\sqrt{X_i})$ para $i = 1, \ldots, n$.

Gerando de outras distribuições

Considere a variável aleatória X com suporte $\{x_1, \ldots, x_k\}$ com probabilidades $p_1, \ldots, p_k$ tal que $\sum_{i=1}^{k} p_i = 1$.

Podemos dividir o intervalo [0,1] em k intervalos $I_1, \ldots, I_k$ de forma que $I_j = (F_{j-1}, F_j]$, em que $F_0 = 0$ e $F_j = \sum_{i=1}^{j} p_i$.

Geramos u_i da $\mathcal{U}(0,1)$ e verificamos em qual intervalo ele caiu. Se u_i em I_j, então dizemos que $X = x_j$ e assim por diante.

Por que esse gerador gera da distribuição de X acima?

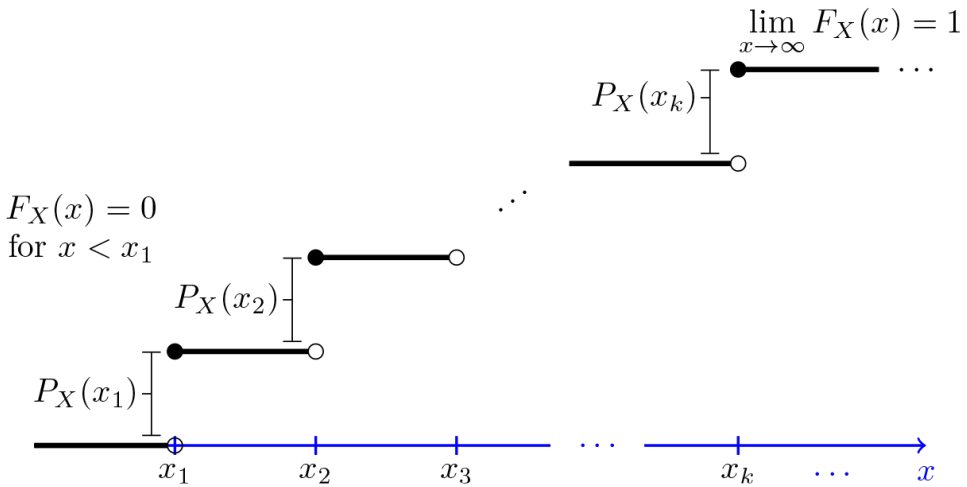


Figura 1: Função de distribuição acumulada.

Exemplo

O objetivo é gerar uma amostra da distribuição de Bernoulli.

Distribuição: $\Pr(X = 1) = p$ e $\Pr(X = 0) = 1 - p$ com $0 < p < 1$.

Notação: $X \sim \mathcal{BR}(p)$.

Algoritmo

1. *Dividir o intervalo (0,1) em 2 partes:* $I_1 = (0, p]$ *e* $I_2 = (p, 1]$.
2. *Gerar u da* $\mathcal{U}(0, 1)$.
3. *Se* $u \in I_1$, *então* $x = 1$. *Caso contrário,* $x = 0$.
4. *Repita n vezes os Passos 1, 2 e 3.*

Ver arquivo exemplo_09.r

Exemplo

O objetivo é gerar uma amostra da distribuição uniforme discreta.

Distribuição: $\Pr(X = x) = \frac{1}{N}$ para $x = 1, 2, \ldots, N$, com $N \in \mathbb{Z}_+$.

Algoritmo

1. *Dividir o intervalo (0,1) em N partes:*
   $I_1 = (0, 1/N], I_2 = (1/N, 2/N], \ldots, I_N = ((N-1)/N, 1]$.
2. *Gerar u da* $\mathcal{U}(0, 1)$.
3. *Se* $u \in I_j$*, então x recebe o valor j.*
4. *Repita n vezes os Passos 1, 2 e 3.*

Ver arquivo exemplo_10.r

Método da transformação inversa

Na representação da estrutura de uma variável aleatória podemos representar a tripla $(\Omega, \mathcal{A}, P)$ por $([0,1], \mathcal{B}, \mathcal{U}(0,1))$.

Isto é, podemos igualar a variabilidade de $\omega \in \Omega$ usando a uniforme no intervalo (0,1) através de uma transformação.

Dessa forma, X são funções que levam do (0,1) em D_x.

Esse método para gerar valores da variável aleatória X é conhecido como método da transformação inversa.

Lema
Se $U \sim \mathcal{U}(0,1)$, então a variável $F^{-1}(U)$ tem função de distribuição acumulada F.

Exercício: prove este lema (pelo menos para variáveis aleatórias contínuas).

Então, basta gerar U da uniforme e fazer $X = F^{-1}(U)$ para obter X com distribuição acumulada F.

Podemos obter outras distribuições como uma transformação determinística de amostras da uniforme.

O `R` utiliza o método da inversa (via `qnorm`) para gerar de uma normal [1].

Exemplo

O objetivo é gerar uma amostra da distribuição exponencial.

A relação $X = -\log(U)$ pode ser utilizada para se obter uma amostra da distribuição exponencial com parâmetro 1.

Por que? Como encontrar essa transformação?

Como generalizar para uma exponencial com parâmetro λ (média $1/\lambda$)?

Ver arquivo exemplo_11.r

[1] Ver `norm.kind` em `help("RNG")`

Exemplo

O objetivo é gerar uma amostra da distribuição gama.

Como utilizar a uniforme, $\mathcal{U}(0,1)$, para gerar da $\mathcal{G}(\kappa,\beta)$, distribuição gama, em que κ é um inteiro positivo e $\beta > 0$?

Algoritmo

1. *Gerar* $u_1, \ldots, u_\kappa$ *da* $\mathcal{U}(0,1)$.
2. *Definir* $y_i = -\frac{1}{\beta}\log(u_i)$, *para* $i = 1, \ldots, \kappa$.
3. *Faça* $x = y_1 + \cdots + y_\kappa$.
4. *Repita n vezes os Passos 1, 2 e 3.*

Ver arquivo exemplo_12.r

Geração de variáveis aleatórias usando misturas

Podemos usar uma técnica bivariada para simplificar a geração de variáveis.

Motivação: sabemos que $f(x_1, x_2) = f(x_1|x_2)f(x_2)$. Suponha que seja mais fácil gerar de $(X_1|X_2 = x_2)$ e de X_2 do que gerar diretamente de X_1.

Então, (X_1, X_2) pode ser gerado em dois passos:

Passo 1: Gerar x_2.

Passo 2: Gerar $(x_1|x_2)$.

Nesse algoritmo, temos disponível a distribuição conjunta (X_1, X_2) e a distribuição marginal de X_1.

Exemplo

O objetivo é gerar uma amostra da variável aleatória X tal que

$$f_X(x) = \omega\phi(x|\mu_1, \sigma_1^2) + (1-\omega)\phi(x|\mu_2, \sigma_2^2),$$

em que $0 < \omega < 1$ e $\phi(.|a, b^2)$ a função de densidade de probabilidade de uma normal com média a e variância b^2.

Seja $Y \sim \mathcal{BR}(\omega)$. Então, podemos dizer que

$$f_{X|Y}(x|y=1) = \phi(x|\mu_1, \sigma_1^2) \quad \text{e} \quad f_{X|Y}(x|y=0) = \phi(x|\mu_2, \sigma_2^2).$$

Consequentemente, uma maneira de obter uma amostra de X é gerar valores de Y e dado o valor de Y gerar valores de X.

Algoritmo

Para $i = 1, \ldots, n$, faça:

1. *Gerar y_i de $\mathcal{BR}(\omega)$;*
2. *Se $y_i = 1$ gerar x_i de $\mathcal{N}(\mu_1, \sigma_1^2)$. Se $y_i = 0$ gerar x_i de $\mathcal{N}(\mu_2, \sigma_2^2)$.*

O algoritmo gera uma amostra $(x_1, \ldots, x_n)$ de $f_X(x)$.

Ver arquivo exemplo_13.r

Exemplo

O objetivo é gerar uma amostra da distribuição *t*-Student.

Se $X|\lambda \sim \mathcal{N}(\mu, \sigma^2/\lambda)$ e $\lambda \sim \mathcal{G}(\nu/2, \nu/2)$, então $X \sim t_\nu(\mu, \sigma^2)$.

Algoritmo

Para $i = 1, \ldots, n$, *faça:*

1. *Gerar* $\lambda_i \sim \mathcal{G}(\nu/2, \nu/2)$;
2. *Gerar* $x_i \sim \mathcal{N}(\mu, \sigma^2/\lambda_i)$.

O algoritmo gera uma amostra $(x_1, \ldots, x_n)$ de $X \sim t_\nu(\mu, \sigma^2)$.

Ver arquivo exemplo_14.r

Método da rejeição

O objetivo é obter uma amostra de X com função de densidade de probabilidade $f(x)$.

Suponha que gerar uma amostra de X através de $f(x)$ seja difícil, mas obter uma amostra de uma $g(x)$ parecida com $f(x)$ seja fácil.

Dizemos que $g(x)$ é um envelope para $f(x)$ se existe uma constante $C \geqslant 1$ tal que $0 \leqslant f(x) \leqslant Cg(x)$ para todo x.

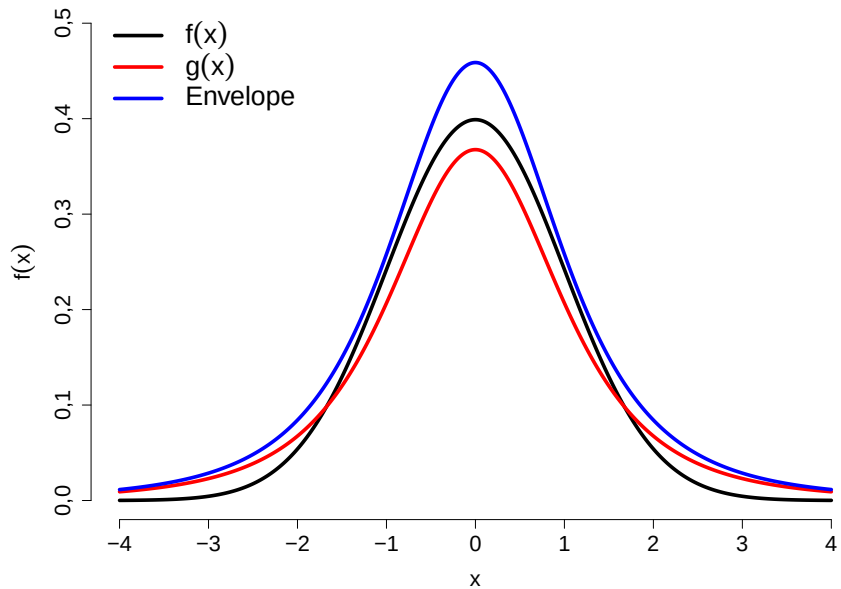


Figura 2: Exemplo do método de rejeição.

Razão de aceitação

Propomos um valor de x gerando da distribuição $g(x)$.

A aceitação desse valor está relacionada com a razão

$$A = \frac{f(x)}{Cg(x)} \quad \text{com} \quad g(x) \neq 0 \quad \text{para todo } x.$$

Pela definição do envelope, temos que $0 \leqslant A \leqslant 1$.

Valores pequenos de A indicam que o valor proposto tem densidade pequena em f e grande em Cg.

Valores de A próximos de 1 indicam que o valor proposto tem densidade alta em f e Cg. Isto é, as duas são parecidas.

Algoritmo

1. *Gere um valor proposto $x^\star$ de $g(x)$;*
2. *Aceite o valor proposto $x^\star$ com probabilidade*

$$A = \frac{f(x^\star)}{Cg(x^\star)}.$$

3. *Repita os Passos 1 e 2 até obter uma amostra de tamanho n.*

O algoritmo gera uma amostra $(x_1, \ldots, x_n)$ de $f(x)$.

Exemplo

O objetivo é gerar uma amostra da normal padrão, $\mathcal{N}(0,1)$, utilizando as seguintes propostas:

(a) $\mathcal{U}(-10, 10)$; e

(b) Cauchy padrão.

No caso da $\mathcal{U}(-10, 10)$, temos

$$C = \frac{(2\pi)^{-1/2} \exp(-x^2/2)}{1/20}.$$

O máximo dessa razão é $C = 7,98$ em $x = 0$. Então,

$$Cg(x) = \frac{7,98}{20} I_{(-10,10)}(x) \approx 0,4 I_{(-10,10)}(x),$$

cobre $f(x)$.

No caso da Cauchy padrão, temos

$$C = \frac{(2\pi)^{-1/2} \exp(-x^2/2)}{[\pi(1 + x^2)]^{-1}}.$$

O máximo dessa razão é $C = 1,52$ em $x = -1$ e $x = 1$. Então, $1,52[\pi(1 + x^2)]^{-1}$ cobre $f(x)$.

Ver arquivo exemplo_15.r

Notas sobre o método da rejeição

A distribuição proposta deve ser fácil de gerar valores.

A função de densidade de probabilidade da proposta deve ser fácil de avaliar.

As funções de densidade de probabilidade f e g devem ser similares para resultar em uma alta probabilidade de aceitação e um algoritmo eficiente.

Dado que $C \geqslant 1$, podemos calcular a probabilidade de aceitação:

$$\begin{aligned} \Pr\left(U \leqslant \frac{f(X)}{Cg(X)}\right) &= \int \Pr\left(U \leqslant \frac{f(s)}{Cg(s)} \middle| X = s\right) g(s)ds \\ &= \int \frac{f(s)}{Cg(s)} g(s)ds = \frac{1}{C}\int f(s)ds = \frac{1}{C}. \end{aligned}$$

Temos que C mais próximo de 1 indica que f e g são próximas, levando a altas probabilidades de aceitação. Enquanto que C grande indica que f e g são distantes, levando a baixas probabilidades de aceitação.

No caso da **proposta uniforme**, a probabilidade de aceitação é **1/7,9788=0,125** e na **proposta Cauchy**, a probabilidade de aceitação é **1/1,5203=0,658**.

Observação importante:

Se só conhecemos a função de densidade de interesse exceto por uma constante de integração

$$f(x) \propto \widetilde{f}(x)$$

nada muda no algoritmo. Basta substituir f por $\widetilde{f}$. Agora, a constante C é tal que

$$\widetilde{f}(x) \leqslant Cg(x).$$

Exemplo

O objetivo é gerar uma amostra de $f(x) \propto \exp\left(-\frac{x^4}{2}\right)$ utilizando como proposta a distribuição normal padrão.

Ver arquivo exemplo_16.r

Exemplo

O objetivo é gerar uma amostra da distribuição normal truncada entre 0 e infinito (conhecida como *half-normal*).

Considere como proposta a distribuição normal padrão.

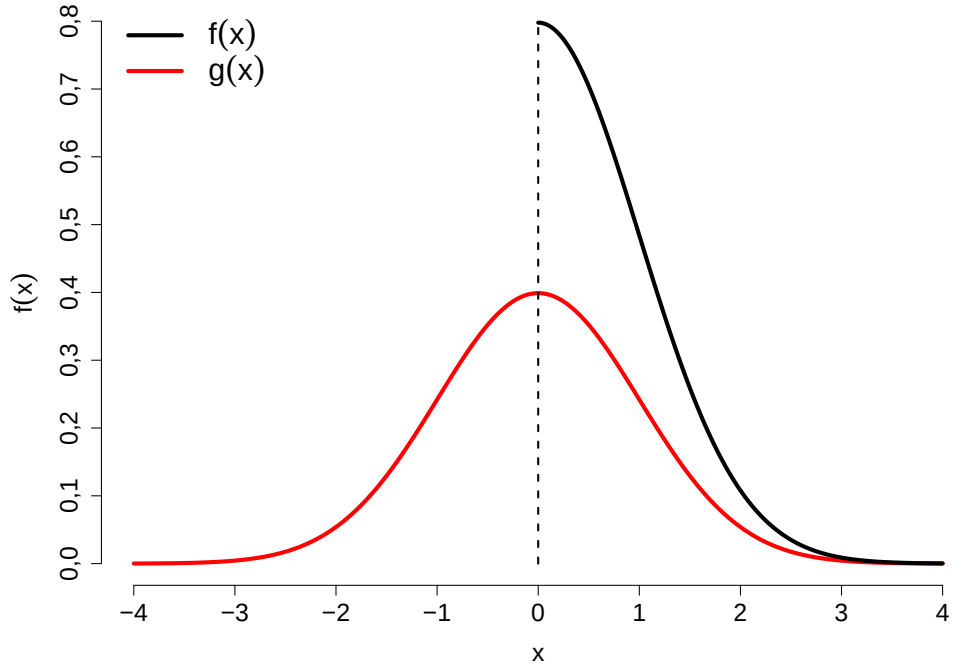


Figura 3: Exemplo do método de rejeição aplicado à normal truncada.

A função de densidade de probabilidade objetivo é

$$f(x) = 2 \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) I_{[0,\infty)}(x) = 2\phi(x) I_{[0,\infty)}(x).$$

No caso da proposta normal padrão com densidade $\phi(x)$, temos

$$C = \frac{2 \times (2\pi)^{-1/2} \exp\left(-x^2/2\right) I_{[0,\infty)}(x)}{(2\pi)^{-1/2} \exp\left(-x^2/2\right)} = 2 I_{[0,\infty)}(x).$$

O máximo dessa razão é $C = 2$ em $x \geqslant 0$. Então, $2\phi(x)$ cobre $f(x)$.

A probabilidade de aceitação é $1/C = 1/2$.

Algoritmo

1. *Gere um valor proposto $x^\star$ de $\mathcal{N}(0,1)$;*
2. *Aceite o valor proposto $x^\star$ com probabilidade*

$$\frac{2\phi(x)}{2\phi(x)} I_{[0,\infty)}(x) = I_{[0,\infty)}(x).$$

3. *Repita os Passos 1 e 2 até obter uma amostra de tamanho n.*

Exemplo

O objetivo é gerar uma amostra da distribuição normal truncada em $(0, \infty)$.

Considere a proposta $\mathcal{G}(1,1) \equiv \mathcal{E}(1)$. Nesse caso, a probabilidade de aceitação é 0,76 ($C = 1,315$ em $x = 1$).

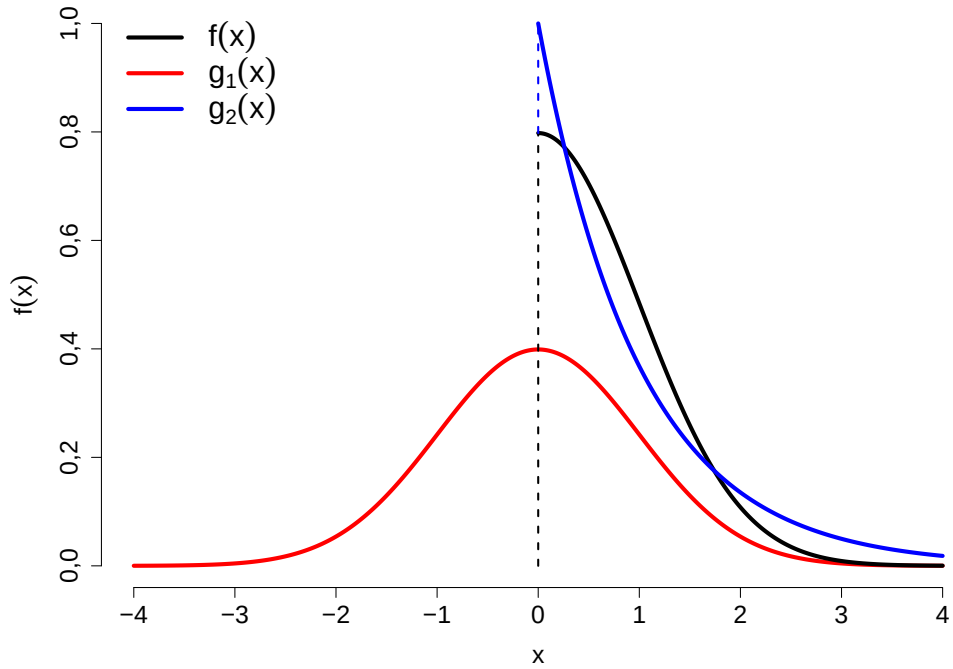


Figura 4: Exemplo do método de rejeição aplicado à normal truncada.

Ver arquivo exemplo_17.r

Gerando da distribuição normal

- Teorema central do limite
- Método da rejeição com proposta exponencial dupla
- Método de Box-Muller

Exemplo

O objetivo é gerar uma amostra da distribuição normal padrão utilizando o **teorema central do limite** (TCL).

Considere uma amostra $(U_1, \ldots, U_n)$ i.i.d. $\mathcal{U}(0,1)$. Pelo TCL, sabemos que para n grande,

$$X = \frac{\sqrt{n}(\overline{U} - 1/2)}{1/\sqrt{12}} \approx \mathcal{N}(0,1).$$

Essa aproximação é razoável para n igual ou maior que 12.

Mas esse algoritmo não é eficiente, pois para gerar um único valor da $\mathcal{N}(0,1)$ é preciso gerar pelo menos 12 valores da $\mathcal{U}(0,1)$.

Ver arquivo exemplo_18.r

Exemplo

O objetivo é gerar uma amostra da distribuição normal padrão pelo **método da rejeição**.

Considere uma proposta Laplace, com $\tau > 0$, dada por

$$g(x|\tau) = \frac{1}{2\tau} \exp\left(-\frac{1}{\tau}|x|\right), \quad x \in \mathbb{R}.$$

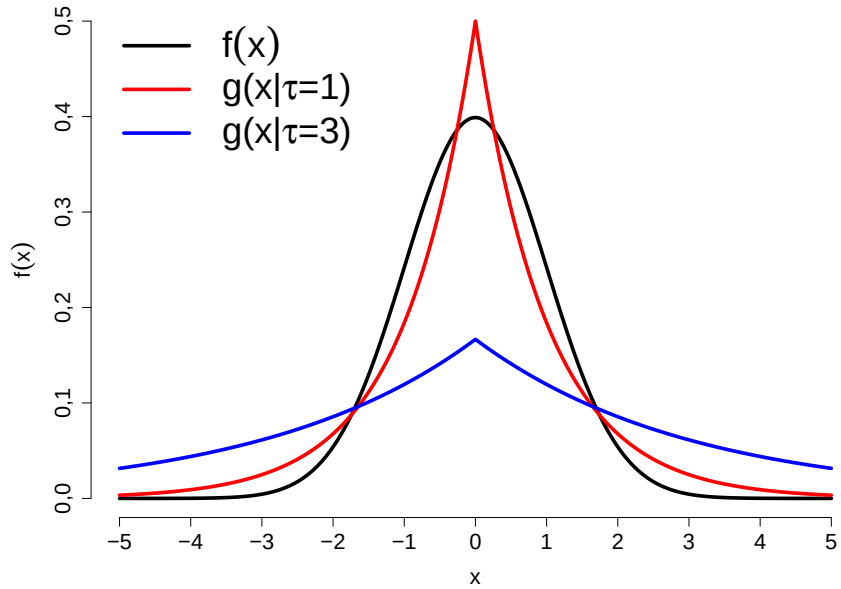


Figura 5: Gerar da normal padrão pelo método de rejeição com proposta Laplace.

A razão entre as densidade f e g é

$$C(\tau) = \frac{f(x)}{g(x)} = \tau\sqrt{\frac{2}{\pi}}\exp\left(-\frac{x^2}{2} + \frac{|x|}{\tau}\right), \quad x \in \mathbb{R}, \quad \tau > 0.$$

O máximo dessa razão é obtida para $x = -1/\tau$ e $x = 1/\tau$ e é igual a

$$C(\tau) = \tau\sqrt{\frac{2}{\pi}}\exp\left(\frac{1}{2\tau^2}\right).$$

Lembrando que a constante $C(\tau)$ deve ser a menor possível para uma alta probabilidade de aceitação.

Então, devemos minimizar $C(\tau)$ com respeito a τ. Para isto, $\tau = 1$.

Segue-se que $C(1) = 1,315$ e a probabilidade de aceitação $1/C(1) = 0,76$.

Portanto, nossa proposta é uma Laplace(1).

Ver arquivo exemplo_19.r

Exemplo

O objetivo é gerar uma amostra da distribuição normal padrão pelo **método Box-Muller**.

Sejam $U_i \sim \mathcal{U}(0,1)$, $i = 1,2$ e independentes. Defina as variáveis aleatórias X_1 e X_2 tais que

$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)}\cos(2\pi U_2), \quad \text{e} \\ X_2 &= \sqrt{-2\log(U_1)}\text{sen}(2\pi U_2). \end{aligned}$$

Então, $X_i \sim \mathcal{N}(0,1)$, $i = 1,2$ e independentes.

Por transformação polar podemos verificar a relação acima.

Note que no algoritmo anterior de rejeição usando proposta Laplace em 100 valores da $\mathcal{U}(0,1)$ gerados esperamos gerar 76 valores da $\mathcal{N}(0,1)$.

Enquanto nesse algoritmo, em 100 valores da $\mathcal{U}(0,1)$ gerados iremos gerar 100 valores da $\mathcal{N}(0,1)$.

Ver arquivo exemplo_20.r

Exemplo

O objetivo é gerar uma amostra da **distribuição normal multivariada** com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$ a partir de variáveis aleatórias $\mathcal{N}(0,1)$ independentes.

Considere $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ um vetor de variáveis aleatórias independentes e identicamente distribuídas da $\mathcal{N}(0,1)$.

Sabemos que $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{AZ} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ em que $\boldsymbol{\Sigma} = \boldsymbol{AA}^\top$.

É possível utilizar a decomposição de Cholesky, $\boldsymbol{\Sigma} = \boldsymbol{U}^\top \boldsymbol{U}$, com $\boldsymbol{A} = \boldsymbol{U}^\top$.

É possível utilizar a decomposição em valores singulares (SVD), $\boldsymbol{\Sigma} = \boldsymbol{VDV}^T$, com $\boldsymbol{A} = \boldsymbol{VD}^{1/2}$.

Contudo, não é possível com a decomposição $\boldsymbol{LU}$.

Ver arquivo exemplo_21.r

Método do envelope

Esse método é parecido com o método de rejeição com um passo extra que diminui o custo computacional no caso em que avaliar a densidade objetivo é difícil.

A ideia do método consiste em encontrar uma constante C tal que Cg cubra a f e escolher uma função h que limite f por baixo.

Se $h(x) \leqslant f(x)$ para todo x, então podemos usar a seguinte relação para aceitar x:

$$u \leqslant \frac{h(x)}{Cg(x)} \leqslant \frac{f(x)}{Cg(x)} \Rightarrow u \leqslant \frac{f(x)}{Cg(x)}.$$

Não é preciso avaliar $f(x)$, basta aceitar o valor proposto.

Exemplo

O objetivo é gerar uma amostra da distribuição $\mathcal{N}(0,1)$ usando como proposta a $\mathcal{U}(-10;\ 10)$.

Utilizaremos um limitante inferior para avaliar a função de densidade da $\mathcal{N}(0,1)$ um número de vezes menor.

Para isto considere a relação $(1 - x^2/2) \leqslant \exp(-x^2/2)$.

O método da rejeição terá um passo extra que testa se

$$u \leqslant \frac{(2\pi)^{-1/2}(1 - x^2/2)}{Cg(x)} = (1 - x^2/2) \quad \text{pois} \quad C = \frac{(2\pi)^{-1/2}}{1/20}.$$

Se a relação for verificada, então podemos aceitar o valor proposto de $g(x)$.

Ver arquivo exemplo_22.r

Reamostragem ponderada

No método de rejeição vimos que é crucial encontrar a constante C para determinar o envelope.

O método da reamostragem ponderada usa a ideia do método de rejeição sem precisar encontrar essa constante.

Desvantagem: produz valores com distribuição aproximada.

Ao invés de gerar diretamente da densidade objetivo $f(x)$, geraremos de uma aproximação discreta dessa distribuição.

Para isso, assim como no método de rejeição, geraremos uma amostra de uma densidade proposta $g(x)$.

Na distribuição discreta as probabilidades dos valores $x_1, x_2, \ldots, x_n$ são tais que

$$\Pr(X = x_i) \propto \frac{f(x_i)}{g(x_i)}.$$

Algoritmo

1. *Considere uma amostra* $(x_1, \ldots, x_n)$ *obtida da função de densidade de probabilidade proposta* $g(x)$*.*
2. *Calcule os pesos*

$$\omega_i = \frac{f(x_i)}{g(x_i)} \quad i = 1, \ldots, n,$$

   e depois as probabilidades

$$\pi_i = \frac{\omega_i}{\sum_{j=1}^{n} \omega_j} \quad i = 1, \ldots, n.$$

3. *Uma segunda amostra* $(x_1^\star, \ldots, x_m^\star)$ *é obtida da distribuição discreta* $\Pr(X = x_i) = \pi_i$*,* $i = 1, \ldots, n$*.*

A reamostra $(x_1^\star, \ldots, x_m^\star)$ tem distribuição aproximada $f(x)$.

Note que o tamanho da reamostra é qualquer. Por exemplo, podemos ter $m > n$ e nesse caso temos amostragem com reposição da distribuição discreta.

Observe que não é preciso conhecer $f(x)$ completamente (só o núcleo).

Exemplo

O objetivo é gerar uma amostra da variável aleatória X tal que

$$f(x) = \omega\phi(x|\mu_1, \sigma_1^2) + (1-\omega)\phi(x|\mu_2, \sigma_2^2),$$

em que $0 < \omega < 1$ e $\phi(.|a, b^2)$ é a função de densidade de probabilidade de uma normal com média a e variância b^2.

Sabemos que

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}_Y(\mathrm{E}(X|Y)) = \omega\mu_1 + (1-\omega)\mu_2 = \mu \\ \mathrm{Var}(X) &= \mathrm{E}_Y(\mathrm{Var}(X|Y)) + \mathrm{Var}_Y(\mathrm{E}(X|Y)) \\ &= [\omega\sigma_1^2 + (1-\omega)\sigma_2^2] + [\omega(\mu_1-\mu)^2 + (1-\omega)(\mu_2-\mu)^2]. \end{aligned}$$

Ver arquivo exemplo_23.r

Exemplo

O objetivo é gerar uma amostra da variável aleatória X tal que

$$f(x) \propto \exp(-x^2/2)\{[\mathrm{sen}(6x)]^2 + 3[\cos(x)]^2[\mathrm{sen}(4x)]^2 + 1\}, \quad x \in \mathbb{R}.$$

Ver arquivo exemplo_24.r

Qualidade da aproximação

Podemos mostrar que para $n \to \infty$ temos $F_\omega(a) \to F_X(a)$ em que

$$F_\omega(a) = \Pr(X \leqslant a) = \sum_{j:x_j \leqslant a} I_{(-\infty,a]}(x_j)\pi_j = \frac{\sum_{j=1}^n I_{(-\infty,a]}(x_j)f(x_j)/g(x_j)}{\sum_{k=1}^n f(x_k)/g(x_k)}.$$

Quando $n \to \infty$, temos vários x_j gerados com frequência de $g(x_j)$. Então, podemos aproximar

$$\begin{aligned} F_\omega(a) &= \sum_{j:x_j \leqslant a} I_{(-\infty,a]}(x_j)\pi_j = \frac{\sum_{j=1}^n I_{(-\infty,a]}(x_j)f(x_j)/g(x_j)}{\sum_{k=1}^n f(x_k)/g(x_k)} \\ &\approx \frac{\int [I_{(-\infty,a]}(x)f(x)/g(x)]g(x)dx}{\int [f(x)/g(x)]g(x)dx} = \frac{\int I_{(-\infty,a]}(x)f(x)dx}{\int f(x)dx} = F_X(a). \end{aligned}$$

Dessa forma, teremos uma boa aproximação para n grande.

Integração numérica

Se a integração analítica não é possível ou é complicada, podemos aproximá-la usando métodos numéricos.

A integração numérica muitas vezes é associada a quadratura - que se refere a encontrar quadrados cuja área seja a mesma sob a curva de interesse.

No caso multivariado, o problema é conhecido como integração múltipla ou cubatura.

É importante no uso de métodos bayesianos e bayesianos empíricos, no cálculo da função de verossimilhança e no cálculo de distribuições a posteriori (momentos, quantis, etc).

Métodos de integração numérica

Métodos determinísticos

- Regra de Riemann
- Regra do trapézio; e
- Regra de Simpson.

Métodos de simulação estocástica

- Método de Monte Carlo simples; e
- Método de Monte Carlo via função de importância.

Métodos assintóticos

- Aproximação Normal; e
- Laplace.

Exemplo

Utilizaremos o cálculo do valor esperado de uma variável aleatória qui-quadrada com ν graus de liberdade para diversos métodos a serem apresentados.

Suponha $X \sim \chi^2_\nu$ e desejamos $\mathrm{E}(X)$, isto é,

$$\mathrm{E}(X) = \int_0^\infty x f(x) dx = \int_0^\infty \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2} \exp(-x/2) dx = \nu.$$

A ideia é resolver a integral acima por métodos numéricos diferentes e comparar o resultado com ν.

Regra de Riemann

A integral de Riemann pode ser definida por

$$\int_a^b f(x)dx = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} \Delta x.f(x_i), \quad \Delta x = \frac{b-a}{n} \quad \text{e} \quad x_i = a + \Delta x.i.$$

Ideia da regra de Riemann: a integral pode ser aproximada por

$$\int_a^b f(x)dx \simeq (b-a)f(b).$$

Usamos o método para vários subintervalos da região de interesse.

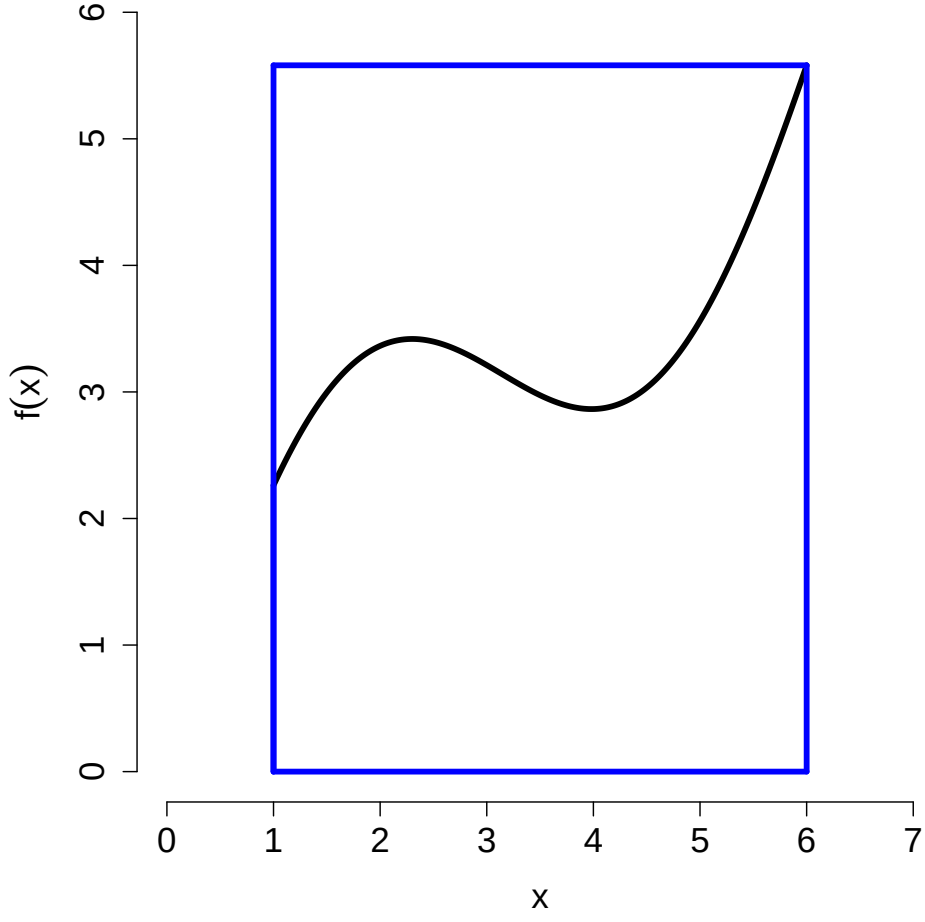


Figura 6: Integral de uma função pela regra de Riemann.

Ver arquivo exemplo_25.r

Regra do trapézio

Considere o problema de resolver a integral $\int_a^b f(x)dx$.

Ideia da regra do trapézio: aproximar a integral sob a curva pela área do trapézio formado por $(a, 0)$, $(b, 0)$, $(a, f(a))$ e $(b, f(b))$. A integral pode ser aproximada por

$$\int_a^b f(x)dx \simeq \frac{b-a}{2}[f(a) + f(b)].$$

Usamos o método para vários subintervalos da região de interesse.

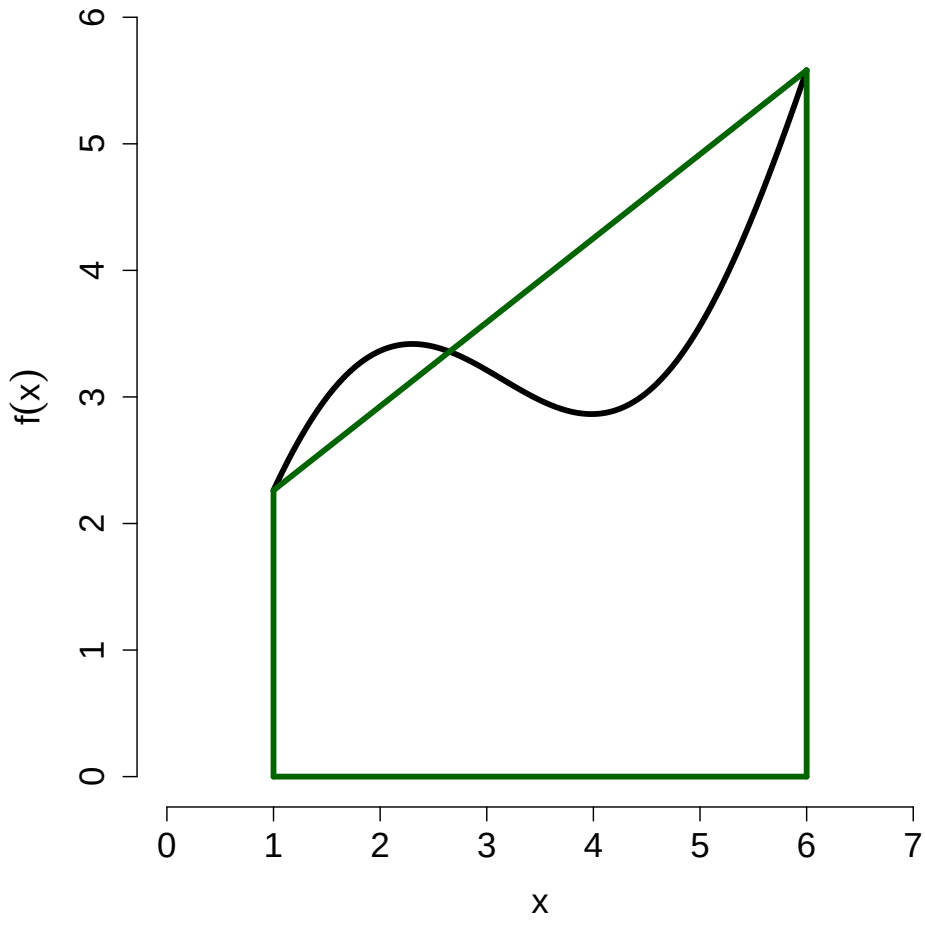


Figura 7: Integral de uma função pela regra do trapézio.

Ver arquivo exemplo_26.r

Regra de Simpson

Ideia da regra de Simpson: aproximar a integral sob a curva pela área sob o polinômio de segunda ordem nos pontos a, $(a+b)/2$ e b. A integral pode ser aproximada por

$$\int_a^b f(x)dx \simeq \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].$$

Se puder, por favor veja "Explanation of Simpson's rule" (em inglês).

Usamos o método para vários subintervalos da região de interesse.

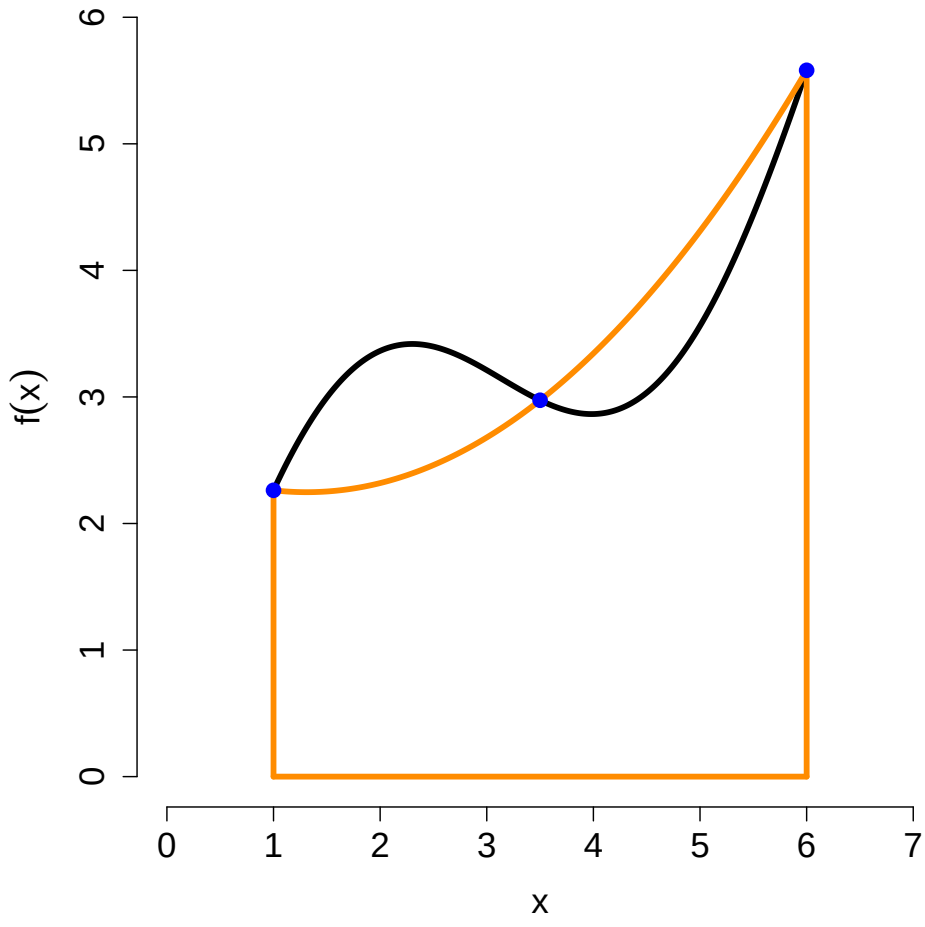


Figura 8: Integral de uma função pela regra de Simpson.

Ver arquivo exemplo_27.r

Métodos de Monte Carlo

Seja $(X_1, \ldots, X_n)$ variáveis aleatórias independentes e identicamente distribuídas de $X \sim \mathcal{N}(\mu, \sigma^2)$ e seja

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_x}. \tag{2}$$

Qual é a distribuição de T?

Mais de 100 anos atrás Gossett usou um experimento de Monte Carlo antes dele provar analiticamente que essa estatística tem distribuição t_{n-1}.

Passos do experimento

1. Entrar com os valores de μ, σ^2, n e m (este último é o número de **replicações** ou **réplicas** do experimento);
2. Para $j = 1, \ldots, m$: gerar uma amostra $(x_1, \ldots, x_n)$ de $\mathcal{N}(\mu, \sigma^2)$ e calcular t_j utilizando a Equação (2); e
3. Estimar a densidade de interesse usando a amostra $(t_1, \ldots, t_m)$.

Ver arquivo exemplo_28.r

Integração de Monte Carlo

O método originalmente foi desenvolvido por físicos para resolver integrais.

Queremos integrar a função $h(x)$ no intervalo (a, b). Podemos escrever a integral de interesse:

$$I = \int_a^b h(x)dx = \int_a^b \frac{h(x)}{f(x)} f(x)dx.$$

Dessa forma, o integrando pode ser visto como o valor esperado de $h(x)/f(x)$ sob a densidade $f(x)$ com suporte no intervalo (a, b).

A **integração de Monte Carlo** é dada por:

$$I = \int_a^b h(x)dx = \int_a^b \frac{h(x)}{f(x)} f(x)dx \simeq \frac{1}{n} \sum_{i=1}^{n} \frac{h(x_i)}{f(x_i)},$$

com $(x_1, \ldots, x_n)$ sendo uma amostra de $f(x)$.

Seja

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{h(x_i)}{f(x_i)}.$$

Pela Lei Forte dos Grandes números temos que $\hat{I}$ converge quase certamente para I.

Podemos estimar a variância dessa aproximação por

$$v_n = \frac{1}{n^2} \sum_{i=1}^{n} \left[\frac{h(x_i)}{f(x_i)} - \hat{I} \right]^2.$$

Temos que $\sqrt{v_n}$ **é conhecido como erro de Monte Carlo**.

Além disso, temos que para n grande

$$\frac{\hat{I} - I}{\sqrt{v_n}} \approx \mathcal{N}(0, 1).$$

Exemplo

Queremos integrar $h(x) = [\cos(50x) + \text{sen}(20x)]^2$ no intervalo $(0, 1)$.

Nesse caso, poderíamos resolver analiticamente. O valor exato dessa integral é 0,965.

Utilizando Monte Carlo:

1. Gerar uma amostra $(u_1, \ldots, u_n)$ de $\mathcal{U}(0, 1)$.
2. Calcular $\frac{1}{n} \sum_{i=1}^{n} h(u_i)$.

Ver arquivo exemplo_29.r

Exemplo

Queremos calcular $\Pr(Z \leqslant 1,96)$ para $Z \sim \mathcal{N}(0,1)$:

$$\Pr(Z \leqslant 1,96) = \int_{-\infty}^{1,96} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \mathrm{E}(I_{(-\infty;1,96]}(Z)).$$

Sabemos que essa integral não pode ser resolvida analiticamente.

Utilizando Monte Carlo:

1. Gerar uma amostra $(z_1, \ldots, z_n)$ de $\mathcal{N}(0,1)$.
2. Calcular $\frac{1}{n} \sum_{i=1}^{n} I_{(-\infty;1,96]}(z_i)$.

Ver arquivo exemplo_30.r

Exemplo

E se o objetivo é o contrário? Calcular o quantil de 97,5%?

Queremos calcular q tal que

$$0,975 = \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \mathrm{E}(I_{(-\infty;q]}(Z)).$$

Sabemos que essa integral não pode ser resolvida analiticamente.

Utilizando Monte Carlo:

1. Gerar uma amostra $(z_1, \ldots, z_n)$ de $\mathcal{N}(0,1)$.
2. Calcular $F_e(q) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty;q]}(z_i)$ para vários valores de q.
3. Escolher q tal que $F_e(q) = 0,975$.

Ver arquivo exemplo_31.r

Exemplo

Suponha que queremos calcular a distribuição preditiva no ponto y. Então,

$$f(y) = \int f(y|\theta) f(\theta) d\theta = \mathrm{E}_\theta(f(y|\theta)),$$

pode ser aproximada por

$$\hat{f}(y) \simeq \frac{1}{n} \sum_{i=1}^{n} f(y|\theta_i)$$

com θ_i gerado da distribuição a priori $f(\theta)$.

Algum problema nessa abordagem? E se a priori for vaga?

Suponha $(Y|\theta) \sim \mathcal{N}(\theta, 1)$ e $\theta \sim \mathcal{N}(0, 4)$.

Passos do método:

1. Gerar uma amostra $(\theta_1, \ldots, \theta_n)$ de $\mathcal{N}(0, 4)$.
2. Avaliar a função de verossimilhança em y dado cada θ_i, $i = 1, \ldots, n$.
3. Tomar a média amostral destes valores.

Ver arquivo exemplo_32.r

Integração usando função de importância

O objetivo é calcular: $I = \int h(x)f(x)dx.$

Suponha que existe uma função de densidade de probabilidade $g(x)$ que aproxime bem a $f(x)$. Então,

$$I = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \mathrm{E}_g\left(h(X)\frac{f(X)}{g(X)}\right).$$

Assim, a integração Monte Carlo nos leva a

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} h(x_i)\frac{f(x_i)}{g(x_i)},$$

em que $(x_1, \ldots, x_n)$ é uma amostra de $g(x)$.

Se definirmos $\omega_i = \dfrac{f(x_i)}{g(x_i)}$ para $i = 1, \ldots, n$, então

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} h(x_i)\omega_i.$$

Exemplo

Suponha que queremos calcular $\Pr(Z \geqslant 4,5)$ em que $Z \sim \mathcal{N}(0,1)$.

Monte Carlo simples: gerar z_i de $Z \sim \mathcal{N}(0,1)$ e aproximar

$$\Pr(Z \geqslant 4,5) = \int_{-\infty}^{\infty} I_{[4,5;\infty)}(z)\phi(z)dz \simeq \frac{1}{n}\sum_{i=1}^{n} I_{[4,5;\infty)}(z_i).$$

Problema: como $\Pr(Z \geqslant 4,5)$ é muito pequena, se n é pequeno teremos uma frequência 0 de $z_i \geqslant 4,5$: $\sum_{i=1}^{n} I_{[4,5;\infty)}(z_i) = 0$ tal que $\Pr(Z \geqslant 4,5) \simeq 0$.

Solução: utilizar uma função de importância $g(z)$ com maior probabilidade na cauda.

$$\begin{aligned}
\Pr(Z \geqslant 4,5) &= \int_{-\infty}^{\infty} I_{[4,5;\infty)}(z)\phi(z)dz \\
&= \int_{-\infty}^{\infty} I_{[4,5;\infty)}(z)\frac{\phi(z)}{g(z)}g(z)dz \simeq \frac{1}{n}\sum_{i=1}^{n} I_{[4,5;\infty)}(z_i)\frac{\phi(z_i)}{g(z_i)}.
\end{aligned}$$

Ver arquivo exemplo_33.r

A densidade de importância

A densidade g pode ser qualquer (desde que $g(x) > 0$ no suporte de X) para o estimador $\hat{I}_1$ da integral I convergir.

Podemos comparar as variâncias desse estimador para diferentes escolhas de g (pois algumas escolhas são melhores que outras).

Temos que $\text{Var}(\hat{I}_1) = \text{E}(\hat{I}_1^2) - \left[\text{E}(\hat{I}_1)\right]^2 = \text{E}(\hat{I}_1^2) - I^2$ com

$$\text{E}_g(\hat{I}_1^2) = \text{E}_g\left(\frac{[h(x)]^2[f(x)]^2}{[g(x)]^2}\right) = \text{E}_f\left(\frac{[h(x)]^2 f(x)}{g(x)}\right).$$

Logo, $\text{Var}(\hat{I}_1)$ é finita se $\text{E}_g(\hat{I}_1^2)$ for finita, mas isto ocorre somente se a razão $f(x)/g(x)$ for limitada para todo x.

Duas condições suficientes para termos $\text{Var}(\hat{I}_1)$ finita:

1. $f(x)/g(x) < c$ (limitada) para todo $x \in \mathbb{D}_x$; e
2. $\mathbb{D}_x$ compacto, $f(x) < a$ e $g(x) > \varepsilon$ para todo $x \in \mathbb{D}_x$.

Essas condições são muito restritivas. No primeiro caso, temos as mesmas condições do algoritmo de rejeição.

Iremos considerar um estimador alternativo com variância finita:

$$\hat{I}_2 = \frac{\sum_{i=1}^{n} h(x_i)\omega_i}{\sum_{i=1}^{n} \omega_i} \quad \text{com} \quad \omega_i = \frac{f(x_i)}{g(x_i)}, \quad i = 1, \ldots, n.$$

Note que neste caso estamos substituindo n por $\sum_{i=1}^{n} \omega_i$.

Como $(1/n)\sum_{i=1}^{n} \omega_i \to 1$ quando $n \to \infty$, o estimador $\hat{I}_2 \to I$ pela Lei Forte dos Grandes Números.

Uma outra abordagem para diminuir a variância do estimador para I é baseada na relação (**Rao-Blackwell**)

$$\text{Var}(E(\delta(X)|Y)) \leqslant V(\delta(X)).$$

Defina $\ell(x) = h(x)f(x)/g(x)$. Podemos utilizar a relação

$$I = \int \ell(x)g(x)dx \quad \text{com} \quad g(x) = \int g(x,y)dy,$$

tal que

$$I = \int\int \ell(x)g(x,y)dxdy = \int \left[\int \ell(x)g(x|y)dx\right] g(y)dy.$$

Temos duas opções:

1. Gerar a amostra $(x_1, \ldots, x_n)$ de $g(x)$ e estimar por

$$\hat{I}_1 = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i).$$

2. Gerar a amostra $(y_1, \ldots, y_n)$ de $g(y)$ e estimar I por

$$\hat{I}_3 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(\ell(X)|y_i).$$

A variância do terceiro estimador é menor que do primeiro.

Exemplo

Considere o problema de calcular o valor esperado de $h(x) = \exp(-x^2)$ para $X \sim t_\nu(0,1)$.

Consideraremos o seguinte:

- Gerar uma amostra $(x_1, \ldots, x_n)$ de $t_\nu(0,1)$ e calcular

$$\hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} \exp(-x_i^2);$$

- Gerar uma amostra $(y_1, \ldots, y_n)$ de $\mathcal{G}(\nu/2, \nu/2)$ e calcular

$$\hat{I}_3 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\exp(-X^2)|y_i) = \frac{1}{n}\sum_{i=1}^{n} (2/y_i + 1)^{-1/2}.$$

Lembre-se que $(X|Y = y) \sim \mathcal{N}(0, 1/y)$ tal que $X \sim t_\nu(0,1)$.

Ver arquivo exemplo_34.r

Aproximação assintótica de integrais

Em inferência bayesiana estamos interessados em informações da distribuição a posteriori:

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)f(\theta)}{f(\boldsymbol{y})} \quad \text{com} \quad f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)f(\theta)d\theta,$$

em que $\boldsymbol{y} = (y_1, \ldots, y_n)$ é uma amostra de $(Y|\theta)$.

Estamos interessados, por exemplo, em $\mathrm{E}(g(\theta)|\boldsymbol{y})$ para diferentes funções g.

- Média a posteriori: $g(\theta) = \theta$.
- Segundo momento ordinário a posteriori: $g(\theta) = \theta^2$.
- Mediana a posteriori: $g(\theta) = I_{(-\infty;c]}(\theta)$ tal que $\mathrm{E}(g(\theta)|\boldsymbol{y}) = 0,5$.

Diversos métodos de integração determinísticos e de Monte Carlo podem ser usados nesse contexto.

Veremos a seguir aproximações baseados em teoria assintótica.

Aproximação Normal

Ideia do método: usar a aproximação por séries de Taylor para o log de $f(\theta|\boldsymbol{y})$ em torno de sua moda.

Lembrando

$$f(x) = \sum_{k=0}^{\infty} f^{(k)}(x_0)\frac{(x - x_0)}{k!}.$$

Considere uma aproximação de ordem 2:

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2}.$$

Para a expansão em torno da moda $\theta^\star$ de $\log f(\theta|\boldsymbol{y})$ temos que $f'(\theta^\star|\boldsymbol{y}) = 0$, então

$$\log f(\theta|\boldsymbol{y}) \simeq \log f(\theta^\star|\boldsymbol{y}) + \left[\frac{\partial^2}{\partial\theta^2}\log f(\theta|\boldsymbol{y})\Big|_{\theta=\theta^\star}\right]\frac{(\theta - \theta^\star)^2}{2}.$$

Assim,

$$f(\theta|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2\psi^2}(\theta-\theta^\star)^2\right) \quad \text{com} \quad \psi^2 = -\left[\frac{\partial^2}{\partial\theta^2}\log f(\theta|\boldsymbol{y})\Big|_{\theta=\theta^\star}\right]^{-1}.$$

Isto é,

$$(\theta|\boldsymbol{y}) \approx \mathcal{N}(\theta^\star, \psi^2).$$

O caso multivariado é análogo com

$$(\boldsymbol{\theta}|\boldsymbol{y}) \approx \mathcal{N}_d(\boldsymbol{\theta}^\star, \boldsymbol{\Psi}).$$

e

$$\boldsymbol{\Psi} = -\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f(\boldsymbol{\theta}|\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right]^{-1}.$$

Exemplo

Considere $(y|\lambda) \sim \text{Poi}(\lambda)$ e $\lambda \sim \mathcal{G}(a, b)$. Sabemos que nesse caso, a posteriori é conjugada e dada por:

$$(\lambda|y) \sim \mathcal{G}(a+y, b+1), \quad \text{E}(\lambda|y) = \frac{a+y}{b+1} \quad \text{e} \quad \text{Var}(\lambda|y) = \frac{a+y}{(b+1)^2}.$$

Consideraremos a aproximação normal para $f(\lambda|y)$.

Suponha que $a+y \geqslant 1$. Então, o ponto de máximo $\lambda^\star = \dfrac{a+y-1}{b+1}$.

Além disso, temos

$$\psi^2 = -\left[\frac{\partial^2}{\partial\lambda^2} \log f(\lambda|y)\Big|_{\lambda=\lambda^\star}\right]^{-1} = \left[\frac{(a+y-1)}{\lambda^2}\Big|_{\lambda=\lambda^\star}\right]^{-1} = \frac{a+y-1}{(b+1)^2}.$$

Logo, $(\lambda|y) \approx \mathcal{N}\left(\dfrac{a+y-1}{b+1}, \dfrac{a+y-1}{(b+1)^2}\right)$.

Ver arquivo exemplo_35.r

Mas note que nesse exemplo $\lambda > 0$ e a distribuição normal utilizada na aproximação está definida nos reais.

Uma solução é melhorar a aproximação usando uma transformação de λ,

Considere $\xi = \log(\lambda)$. Então, $\lambda = \exp(\xi)$.

Como $f(\lambda|y) \propto \lambda^{a+y-1} \exp(-(b+1)\lambda)$ podemos obter

$$f(\xi|y) \propto \exp(\xi(a+y) - (b+1)\exp(\xi)).$$

Daí, podemos mostrar que

$$(\xi|y) \approx \mathcal{N}\left(\log\left(\frac{a+y}{b+1}\right), \frac{1}{a+y}\right) \Rightarrow (\lambda|y) \approx \mathcal{LN}\left(\log\left(\frac{a+y}{b+1}\right), \frac{1}{a+y}\right),$$

tal que

$$\mathrm{E}(\lambda|y) \approx \left(\frac{a+y}{b+1}\right) \exp\left(\frac{1}{2(a+y)}\right).$$

Ver arquivo exemplo_36

Método de Laplace

Usado para aproximar a razão de duas integrais.

Em particular, muito útil em inferência bayesiana.

Suponha que estejamos interessados em $\mathrm{E}(h(\theta)|\boldsymbol{y})$ sob a distribuição a posteriori para θ. Sabemos que

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)f(\theta)}{f(\boldsymbol{y})} \quad \text{em que} \quad f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)f(\theta)d\theta$$

é a preditiva de $\boldsymbol{y}$. Então,

$$\mathrm{E}(h(\theta)|\boldsymbol{y}) = \frac{\int h(\theta)f(\boldsymbol{y}|\theta)f(\theta)d\theta}{\int f(\boldsymbol{y}|\theta)f(\theta)d\theta}.$$

Podemos utilizar o método de Laplace para aproximar $\mathrm{E}(h(\theta)|\boldsymbol{y})$.

Esse método é bastante usado porque a preditiva $f(\boldsymbol{y})$ também é usualmente difícil de ser obtida analiticamente.

Usamos a mesma aproximação anterior, porém agora temos numerador e denominador para serem aproximados.

Isso leva a uma aproximação melhor, pois parte dos erros se cancelam.

Suponha que $h(\theta) > 0$, e defina

$$\begin{aligned} L_1(\theta) &= \log(f(\boldsymbol{y}|\theta)) + \log(f(\theta)) \\ L_2(\theta) &= \log(f(\boldsymbol{y}|\theta)) + \log(f(\theta)) + \log(h(\theta)). \end{aligned}$$

tal que

$$\mathrm{E}(h(\theta)|\boldsymbol{y}) = \frac{\int h(\theta) f(\boldsymbol{y}|\theta) f(\theta) d\theta}{\int f(\boldsymbol{y}|\theta) f(\theta) d\theta} = \frac{\int \exp(L_2(\theta)) d\theta}{\int \exp(L_1(\theta)) d\theta}.$$

Utilizaremos expansão de Taylor de segunda ordem para calcular uma aproximação para a integral acima.

Seja $\theta_i^\star$ o valor que maximiza $L_i(\theta)$. Seja $\psi_i^2 = -[L_i''(\theta_i^\star)]^{-1}$, para $i = 1, 2$.

Podemos mostrar que

$$\mathrm{E}(h(\theta)|\boldsymbol{y}) \simeq \exp(L_2(\theta_2^\star) - L_1(\theta_1^\star)) \left(\frac{\psi_2^2}{\psi_1^2}\right)^{1/2}.$$

No caso multivariado,

$$\mathrm{E}(h(\boldsymbol{\theta})|\boldsymbol{y}) \simeq \exp(L_2(\boldsymbol{\theta}_2^\star) - L_1(\boldsymbol{\theta}_1^\star)) \left(\frac{|\boldsymbol{\Psi}_2|}{|\boldsymbol{\Psi}_1|}\right)^{1/2}$$

em que

$$\boldsymbol{\Psi}_i = -\left[\frac{\partial^2 \log L_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right]^{-1}, \quad i = 1, 2.$$

Exemplo

O total de n animais está categorizado em 3 categorias $y = (y_1, y_2, y_3)^\top$ com probabilidades $(0,25(2+\theta);\ 0,5(1-\theta);\ 0,25\theta)$.

Temos a função de verossimilhança

$$f(\boldsymbol{y}|\theta) \propto (2+\theta)^{y_1}(1-\theta)^{y_2}\theta^{y_3}, \quad \theta \in (0,1).$$

Considere priori $\theta \sim \mathcal{U}(0,1)$.

Observou-se $\boldsymbol{y} = (125;\ 38;\ 34)^\top$.

Como encontrar a média e o desvio padrão a posteriori de θ?

Devemos encontrar $\frac{dL_1(\theta)}{d\theta}$, $\frac{d^2L_1(\theta)}{d\theta^2}$, $\frac{dL_2(\theta)}{d\theta}$, $\frac{d^2L_2(\theta)}{d\theta^2}$, $\frac{dL_3(\theta)}{d\theta}$ e $\frac{d^2L_3(\theta)}{d\theta^2}$.

Ver arquivo exemplo_37.r

Otimização estocástica

Considere o problema de encontrar o valor que maximiza uma função de interesse.

Podemos usar os métodos numéricos usuais ou usar simulação estocástica.

Variáveis auxiliares são introduzidas no problema (artificialmente ou não) de forma a facilitar a maximização.

Arrefecimento simulado *(Simulated anneling)*

Método utilizado para maximização ou minimização de funções.

Deseja-se maximizar ou minimizar $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^p$.

Método introduzido por Metropolis et al. (1953).

Ideia: mudar a escala (temperatura) da função objetivo permite mudanças mais rápidas de uma moda para outra na superfície de interesse.

Isto é, a mudança de escala evita que o método fique preso em modas locais.

Maximização de $f(x)$

Dada uma temperatura $T > 0$, definimos uma nova função objetivo dada por

$$g(x) \propto \exp\left(\frac{f(x)}{T}\right), \quad \text{para} \quad x \in \mathbb{R}.$$

O valor que maximiza $g(x)$ também maximiza $f(x)$.

Um valor $x^\star$ é proposto da distribuição $\mathcal{U}(x^{(i-1)} - \Delta, x^{(i-1)} + \Delta)$. Esse valor é aceito com probabilidade α dada por:

$$\alpha = \min\left(1, \frac{g(x^\star)}{g(x^{(i-1)})}\right).$$

A função g é utilizada com o objetivo de diminuir vales e picos, tornando possível passar de uma moda local para uma outra moda.

A função g depende de T, que diz quanto a função deve ser suavizada.

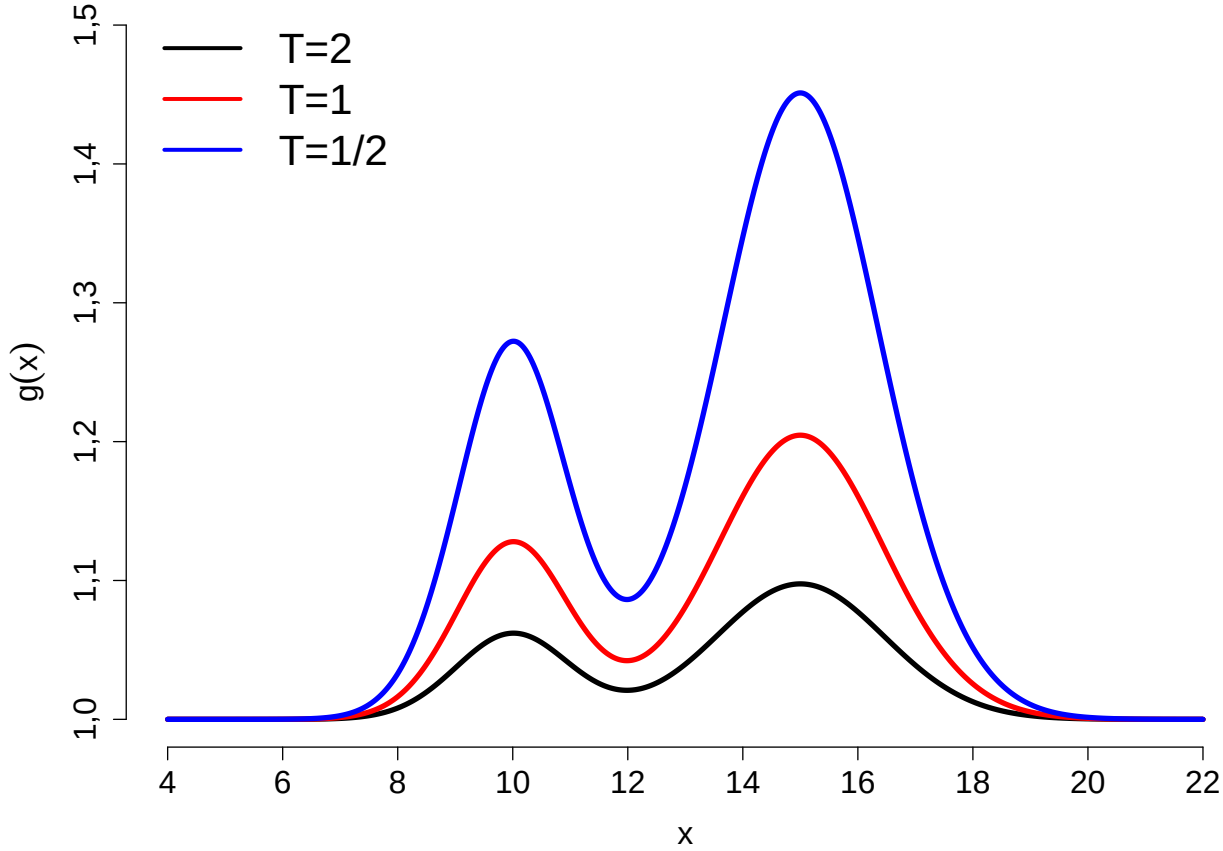


Figura 9: Exemplo da função $g(x)$ para diferentes temperaturas.

- Para $T < 1$ a função se torna mais concentrada e os picos mais altos.
- Para $T > 1$ a função se torna mais vaga e os picos menos acentuados.

Temperatura

A temperatura T é grande no início do algoritmo, de forma a permitir passar de modas locais para moda global e deve ser reduzida aos poucos até que a moda global seja atingida.

As questões que devem ser resolvidas são:

- Definição de T;
- Escolha de Δ; e
- Escolha do esquema de redução da temperatura.

Um reinício pode ser incluído, isto é, o algoritmo é reiniciado partindo do melhor valor até aquele momento para a otimização da função.

Para o caso multivariado basta escolher $(x_1^\star, \ldots, x_p^\star)$ da distribuição uniforme p-variada (hipercubo de dimensão p).

Exemplo

Deseja-se maximizar a função:

$$f(x) = 0,3\phi\left(x - 10\right) + \frac{0,7}{\sqrt{1,5}}\phi\left(\frac{x - 15}{\sqrt{1,5}}\right).$$

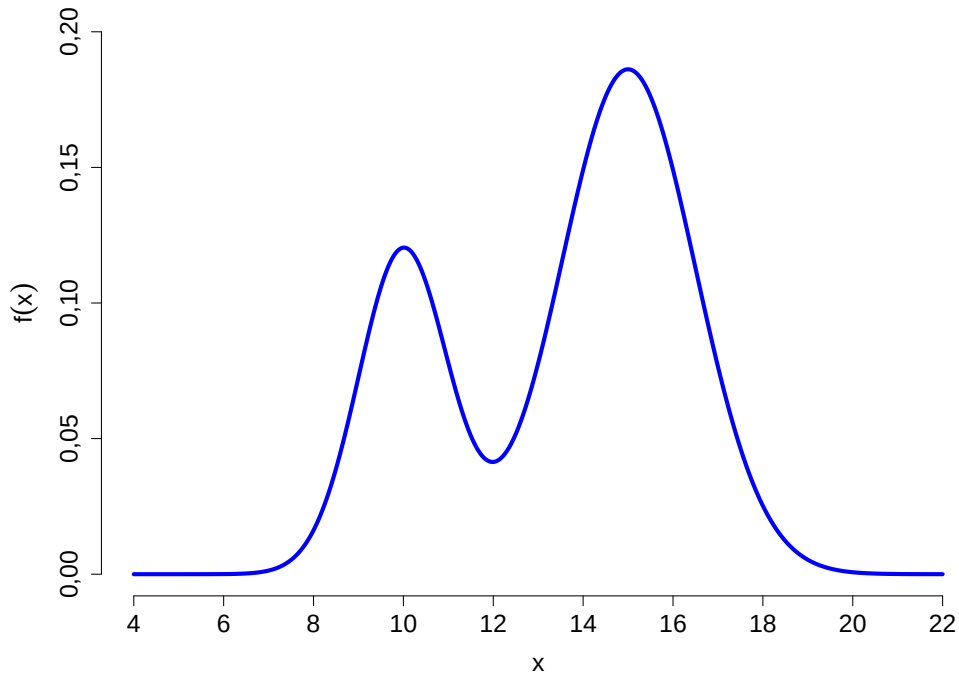


Figura 10: Maximizar a função mistura discreta de normais.

Ver arquivo exemplo_38.r

Exemplo

Deseja-se maximizar a função:

$$f(x) = (\cos(50x) + \text{sen}(20x))^2, \quad x \in (0;\ 0,5).$$

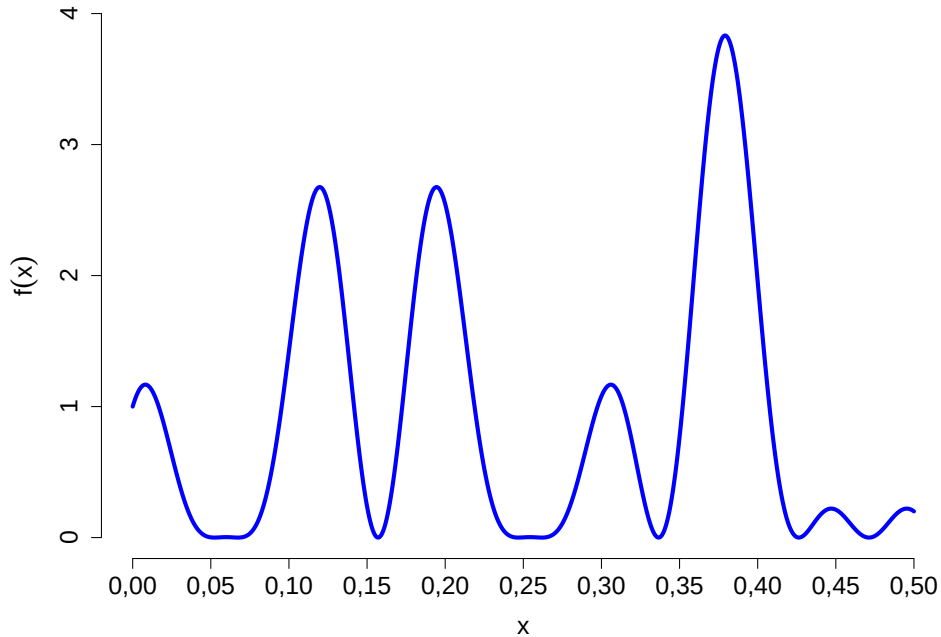


Figura 11: Função $f(x) = (\cos(50x) + \text{sen}(20x))^2$.

Ver arquivo exemplo_39.r

Exemplo

Deseja-se maximizar a seguinte função:

$$\begin{aligned} f(x,y) &= 0,60\exp\{-0,5[(x-2)^2+(y-2)^2]\} \\ &+ 0,35\exp\{-0,5[(x-5)^2+(y-5)^2]\} \\ &+ 0,05\exp\{-0,5[(x-3)^2+(y-6)^2]\}. \end{aligned}$$

Esta função possui um máximo local no ponto $(5,5)$ e o máximo global no ponto $(2,2)$.

Ver arquivo exemplo_40.r

Considere como acerto o algoritmo atingir o ponto $(2,2)$ para $f(2,2) = 0,6000534$, então o critério utilizado foi acerto quando $|\max - f(2,2)| < 0,01$, em que $\max$ é o valor da função no ponto ótimo do algoritmo. Seja k tal que a cada k iterações a temperatura é reduzida em 10% no algoritmo. O ponto inicial foi escolhido aleatoriamente da $\mathcal{U}((1,7) \times (1,7))$.

k	Δ	T inicial	% de acerto
100	0,001	10	4
		10^3	12
		10^7	12
100	0,02	10	96
		10^3	86
		10^7	86
100	1	10	100
		10^3	100
		10^7	100

Note que para $\Delta = 0,001$ os resultados obtidos são muito ruins, para $\Delta = 0,02$ são bastante bons e para $\Delta = 1$ são excelentes.

Para aumento na temperatura inicial, dependendo do Δ utilizado, tanto pode resultar em melhora como em piora das estimativas. Por exemplo, para $\Delta = 0,02$ e $k = 1.000$ quando T é aumentado as estimativas melhoram, mas para $k = 100$ acontece o contrário. Entretanto, há efeito da temperatura.

Algoritmo Esperança-Maximização (EM)

O algoritmo EM pode ser aplicado em modelos de dados faltantes.

Estes modelos têm uma função de verossimilhança que pode ser expressa por

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \int f(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z}.$$

A função de interesse é a marginal de um modelo conjunto para (X, Z).

Exemplo

Considere $(x_1, \ldots, x_n)$ uma amostra aleatória de Poi(τ_i) e $(y_1, \ldots, y_n)$ uma amostra aleatória de Poi$(\beta\tau_i)$, independente da primeira.

A função de verossimilhança para $(\beta, \tau_1, \ldots, \tau_n)^\top$ é dada por

$$\begin{aligned} L(\beta, \boldsymbol{\tau}|\boldsymbol{x}, \boldsymbol{y}) &= \prod_{i=1}^{n} f_X(x_i|\tau_i) f_Y(y_i|\tau_i, \beta) \\ &= \prod_{i=1}^{n} \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}. \end{aligned}$$

Podemos mostrar que

$$\hat{\beta} = \frac{\overline{y}}{\overline{x}}, \quad \text{e} \quad \hat{\tau}_i = \frac{x_i + y_i}{\hat{\beta} + 1}, \quad i = 1, \ldots, n.$$

Suponha que x_1 é um dado faltante. Neste caso, como poderíamos obter $\hat{\tau}_1 = (x_1 + y_1)/(\hat{\beta} + 1)$ e $\hat{\beta} = \overline{y}/\overline{x}$?

Uma opção é ignorar a observação y_1 e considerar os dados $((x_2, y_2), \ldots, (x_n, y_n))$. Contudo, estaríamos descartando dados observados.

Outra opção é considerar a função verossimilhança para dados incompletos ou faltantes,

$$f(y_1, (x_2, y_2), \ldots, (x_n, y_n)|\beta, \tau) = \int f((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)|\beta, \tau) dx_1.$$

Definimos $f(y_1, (x_2, y_2), \ldots, (x_n, y_n)|\beta, \tau)$ como a função de verossimilhança para os dados incompletos.

Nesse problema, temos duas funções de verossimilhanças:

- Dados completos

$$L(\beta, \tau|(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) = f((x_1, y_1), \ldots, (x_n, y_n)|\beta, \tau).$$

- Dados incompletos

$$L(\beta, \tau|y_1, (x_2, y_2), \ldots, (x_n, y_n)) = \int f((x_1, y_1), \ldots, (x_n, y_n)|\beta, \tau) dx_1.$$

Exemplo

Suponha que a variável aleatória X tem distribuição dada por uma mistura tal que

$$f_X(x|\theta) = \theta f_1(x) + (1-\theta) f_2(x),$$

em que f_1 e f_2 são funções de densidade de probabilidades conhecidas.

Para uma amostra aleatória $(X_1, \ldots, X_n)$ com distribuição f_X temos a função de verossimilhança dada por

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} [\theta f_1(x_i) + (1-\theta) f_2(x_i)],$$

Considere introduzir no problema variáveis aleatórias auxiliares $Z_1, Z_2, \ldots, Z_n$ tal que

$$Z_i \sim \text{Ber}(\theta), \quad i = 1, \ldots, n$$

Dessa forma,

$$\begin{aligned} f(x_i|z_i = 1) &= f_1(x_i) \\ f(x_i|z_i = 0) &= f_2(x_i). \end{aligned}$$

Ao integrarmos Z_i, obtemos a função de densidade de probabilidade original para X_i.

Podemos mostrar que a função de verossimilhança completa é dada por

$$L^c(\theta|\boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \left\{ [f_1(x_i)]^{z_i} \times [f_2(x_i)]^{1-z_i} \times \theta^{z_i} \times (1-\theta)^{1-z_i} \right\}.$$

Se conhecêssemos $(z_1, \ldots, z_n)$, então encontraríamos facilmente o estimador de máxima verossimilhança para θ na função de verossimilhança completa.

Algoritmo Esperança-Maximização (EM)

Método iterativo que calcula o estimador de máxima verossimilhança em casos de dados incompletos ou faltantes.

Transformamos o problema de encontrar o máximo de uma função complicada em dois problemas mais simples: esperança (E) e maximização (M).

Suponha que observamos uma amostra aleatória $(X_1, \ldots, X_n)$ da distribuição $f(x|\theta)$.

Considere a função de verossimilhança $L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$.

Queremos encontrar $\hat{\theta} = \text{argmax}_\theta \; L(\theta|\boldsymbol{x})$.

Considere as variáveis auxiliares ou dados faltante $(Z_1, \ldots, Z_n)$ e a distribuição conjunta para (X, Z), $f(x, z|\theta)$.

Seja X a variável aleatória observada e Z a variável aleatória faltante. De forma geral temos:

- Função de verossimilhança para dados completos,

$$L^c(\theta|\boldsymbol{x},\boldsymbol{z}) = f(\boldsymbol{x},\boldsymbol{z}|\theta); \quad \text{e}$$

- Função de verossimilhança para dados incompletos,

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \int f(\boldsymbol{x},\boldsymbol{z}|\theta)d\boldsymbol{z}.$$

Identidade básica do algoritmo

Sabemos que a distribuição condicional de $\boldsymbol{z}$ faltante dado os observáveis e θ é dada por

$$f(\boldsymbol{z}|\boldsymbol{x},\theta) = \frac{f(\boldsymbol{x},\boldsymbol{z}|\theta)}{f(\boldsymbol{x}|\theta)}.$$

Daí,

$$\log f(\boldsymbol{z}|\boldsymbol{x},\theta) = \log f(\boldsymbol{x},\boldsymbol{z}|\theta) - \log f(\boldsymbol{x}|\theta) = \ell^c(\theta|\boldsymbol{x},\boldsymbol{z}) - \ell(\theta|\boldsymbol{x}).$$

em que $\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})$ e $\ell(\theta|\boldsymbol{x})$ são os logaritmos das funções de verossimilhanças completa e incompleta, respectivamente.

Se tomamos a esperança na distribuição de $(\boldsymbol{z}|\boldsymbol{x},\theta)$ para $\theta = \theta_0$ obtemos a relação

$$\ell(\theta|\boldsymbol{x}) = \mathrm{E}(\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})|\boldsymbol{x},\theta_0) - \mathrm{E}(\log(f(\boldsymbol{z}|\boldsymbol{x},\theta))|\boldsymbol{x},\theta_0).$$

Como desejamos maximizar $\ell(\theta|\boldsymbol{x})$, **o segundo termo pode ser ignorado** e maximizamos somente

$$\mathrm{E}(\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})|\theta_0,\boldsymbol{x}).$$

Algoritmo

1. Calcule a esperança $Q(\theta|\boldsymbol{x},\hat{\theta}_{(m)})$ em que

   $$Q(\theta|\boldsymbol{x},\theta_0) = \mathrm{E}(\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})|\theta_0,\boldsymbol{x}),$$

   sendo a esperança calculada na distribuição de $(\boldsymbol{z}|\boldsymbol{x},\theta_0)$.
2. Maximize $Q(\theta|\boldsymbol{x},\hat{\theta}_{(m)})$ em θ:

   $$\hat{\theta}_{(m+1)} = \mathrm{argmax}_\theta \; Q(\theta|\boldsymbol{x},\hat{\theta}_{(m)}).$$

Podemos mostrar que

$$L(\hat{\theta}_{(m+1)}|\boldsymbol{x}) \geqslant L(\theta_{(m)}|\boldsymbol{x}),$$

com igualdade se, e somente se, $Q(\hat{\theta}_{(m+1)}|\boldsymbol{x},\hat{\theta}_{(m)}) = Q(\hat{\theta}_{(m)}|\boldsymbol{x},\hat{\theta}_{(m)})$

Resultado

Mostraremos que o algoritmo obtém uma sequência $\theta_{(j)}$, para $j = 1, 2, \ldots, M$, que converge para o argumento $\theta^\star$ que maximiza $\ell(\theta|\boldsymbol{x})$.

Agora, seja

$$\ell(\theta|\boldsymbol{x}) = \mathrm{E}(\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})|\boldsymbol{x},\theta_0) - \mathrm{E}(\log(f(\boldsymbol{z}|\boldsymbol{x},\theta))|\boldsymbol{x},\theta_0) = Q(\theta|\boldsymbol{x},\theta_0) - W(\theta|\boldsymbol{x},\theta_0),$$

com $Q(\theta|\boldsymbol{x},\theta_0) = \mathrm{E}(\ell^c(\theta|\boldsymbol{x},\boldsymbol{z})|\boldsymbol{x},\theta_0)$ e $W(\theta|\boldsymbol{x},\theta_0) = \mathrm{E}(\log(f(\boldsymbol{z}|\boldsymbol{x},\theta))|\boldsymbol{x},\theta_0)$.

Para mostrar que $\ell(\theta|\boldsymbol{x}) \geqslant \ell(\theta_0|\boldsymbol{x})$, precisamos que

- $Q(\theta|\boldsymbol{x},\theta_0) \geqslant Q(\theta_0|\boldsymbol{x},\theta_0)$. **Isto é verdade pela definição do algoritmo.**
- $W(\theta|\boldsymbol{x},\theta_0) \leqslant W(\theta_0|\boldsymbol{x},\theta_0)$. **Precisamos mostrar que**

$$\begin{aligned} \mathrm{E}(\log(f(\boldsymbol{z}|\boldsymbol{x},\theta))|\boldsymbol{x},\theta_0) &\leqslant \mathrm{E}(\log(f(\boldsymbol{z}|\boldsymbol{x},\theta_0))|\boldsymbol{x},\theta_0) \Leftrightarrow \\ \mathrm{E}\left(\log\left(\frac{f(\boldsymbol{z}|\boldsymbol{x},\theta)}{f(\boldsymbol{z}|\boldsymbol{x},\theta_0)}\right)\middle|\boldsymbol{x},\theta_0\right) &\leqslant 0. \end{aligned}$$

Pela desigualdade de Jensen, temos que

$$\begin{aligned}
\mathrm{E}\left(\log\left(\frac{f(\boldsymbol{z}|\boldsymbol{x},\theta)}{f(\boldsymbol{z}|\boldsymbol{x},\theta_0)}\right)\middle|\boldsymbol{x},\theta_0\right) &\leqslant \log\left(\mathrm{E}\left(\frac{f(\boldsymbol{z}|\boldsymbol{x},\theta)}{f(\boldsymbol{z}|\boldsymbol{x},\theta_0)}\middle|\boldsymbol{x},\theta_0\right)\right) \\
&= \log\left(\int \frac{f(\boldsymbol{z}|\boldsymbol{x},\theta)}{f(\boldsymbol{z}|\boldsymbol{x},\theta_0)} f(\boldsymbol{z}|\boldsymbol{x},\theta_0) d\boldsymbol{z}\right) \\
&= \log(1) \\
&= 0,
\end{aligned}$$

como queríamos.

Exemplo

Considere novamente o modelo Poisson com x_1 como dado faltante.

Podemos considerar o **Algoritmo EM** para encontrar as estimativas de máxima verossimilhança.

Precisamos calcular

$$\mathrm{E}(\ell^c(\beta, \boldsymbol{\tau}|\boldsymbol{x}, \boldsymbol{y})|\beta_0, \boldsymbol{\tau}_0),$$

na distribuição de $X_1|\beta_0, \boldsymbol{\tau}_0$, isto é, em $(X_1|\beta_0, \boldsymbol{\tau}_0) \sim \mathrm{Poi}(\tau_{1,0})$.

Passo E: Calculamos o valor esperado

$$\begin{aligned} Q(\boldsymbol{\tau}, \beta|\tau_{1,0}) &= \mathrm{E}(\ell^c(\beta, \boldsymbol{\tau}|\boldsymbol{x}, \boldsymbol{y})|\beta_0, \boldsymbol{\tau}_0) \\ &= c + \sum_{i=2}^{n}[-\tau_i + x_i \log(\tau_i)] + \sum_{i=1}^{n}[-\beta\tau_i + y_i(\log(\beta) + \log(\tau_i))] \\ &- \tau_1 + \tau_{1,0} \log(\tau_1). \end{aligned}$$

Passo M: Maximizamos a função $Q(\boldsymbol{\tau}, \beta|\tau_{1,0})$ em $(\boldsymbol{\tau}, \beta)$.

Vimos que no caso de dados completos $((x_1, y_1), \ldots, (x_n, y_n))$,

$$\hat{\beta} = \frac{\overline{y}}{\overline{x}}, \quad \text{e} \quad \hat{\tau}_i = \frac{x_i + y_i}{\hat{\beta} + 1}, \quad i = 1, \ldots, n.$$

No caso de x_1 faltante, o algoritmo EM **(Passo M)** é dado por

$$\begin{aligned}
\hat{\beta}_{(m+1)} &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=2}^{n} x_i + \tau_{1,(m)}} \\
\hat{\tau}_{i,(m+1)} &= \frac{x_i + y_i}{\hat{\beta}_{(m+1)} + 1}, \quad i = 2, \ldots, n \\
\hat{\tau}_{1,(m+1)} &= \frac{\hat{\tau}_{1,(m)} + y_1}{\hat{\beta}_{(m+1)} + 1}.
\end{aligned}$$

Ver arquivo exemplo_41.r

Básico de cadeias de Markov

Cadeia de Markov: sequência de variáveis aleatórias que podem ser pensadas como se evoluissem com o tempo.

A evolução no tempo se dá com probabilidade de transição do estado X_n para X_{n+1} dependendo de um núcleo de transição.

Núcleo de transição para $\mathcal{D}_X$ discreto,

$$P_{xy} = \Pr(X_{n+1} = y | X_n = x), \quad x, y \in \mathcal{D}_X.$$

Núcleo de transição para $\mathcal{D}_X$ contínuo. Se $K(.|x)$ é um núcleo de transição, então

$$\Pr(Y \in A|x) = \int_A K(y|x)dy.$$

Consideraremos cadeias de Markov construídas com núcleo de transição $K(.|x)$ que é a função de densidade de probabilidade condicional de $X_{n+1}|X_n$.

Exemplo

Uma sequência de variáveis aleatórias (X_n) é dita um **passeio aleatório** se

$$X_{n+1} = X_n + \varepsilon_n,$$

em que ε_n é gerado independentemente de $X_n, X_{n-1}, \ldots$ Se a distribuição de ε_n é simétrica no zero, dizemos que temos um passeio aleatório simétrico.

- Contínuo: $X_0 = 0$, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$.

Ver arquivo exemplo_42.r

- Exemplo: $X_0 = 0$, $\Pr(\varepsilon_n = -1) = \Pr(\varepsilon_n = 1) = \frac{1}{2}$.

Ver arquivo exemplo_43.r

Cadeia de Markov: dado um núcleo de transição K, uma sequência de variáveis aleatórias $X_0, X_1, \ldots, X_n, \ldots$ é uma cadeia de Markov denotada por (X_n), se, para qualquer k,

$$P(X_{k+1} \in A | x_0, x_1, \ldots, x_k) = P(X_{k+1} \in A | x_k)$$

Exemplo

A sequência $X_0, X_1, \ldots$ gerados no algoritmo **arrefecimento simulado** formam uma cadeia de Markov em que

$$X_{n+1} = \begin{cases} X^\star, & \text{com probabilidade} \quad \min\{1, \exp\{[g(x^\star) - g(x_n)]/T\}\} \\ X_n, & \text{caso contrário.} \end{cases}$$

Se T muda ao longo das iterações, então dizemos que a cadeia é heterogênea no tempo.

Cadeias de Markov que possuem uma **distribuição estacionária** f são tais que se $X_n \sim f$, então $X_{n+1} \sim f$, quando o núcleo de transição K permite movimentos em todo espaço.

Nosso objetivo: usar cadeias de Markov para gerar amostras da distribuição $f(\boldsymbol{x})$, de forma que $f(\boldsymbol{x})$ seja a distribuição estacionária da cadeia. Em aplicações, pode ser uma posteriori: $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Contudo, precisamos construir cadeias com certas propriedades para alcançar esse objetivo. Por exemplo, a cadeia deve ser recorrente, estacionária, reversível e ergótica.

Recorrência

Em um espaço de estados finito $\mathcal{D}_X$, um estado $\omega \in \mathcal{D}_X$ é dito **transiente** se o número médio de visitas a ω é finito. E um estado é dito **recorrente** se o número médio de visitas a ω é infinito.

Uma cadeia é dita **recorrente se o número médio de visitas a um conjunto qualquer A é infinito.**

Exemplo

Considere o modelo autoregressivo de ordem 1, AR(1), dado por

$$X_n = \theta X_{n-1} + \varepsilon_n, \quad |\theta| < 1 \quad \text{e} \quad \varepsilon_n \sim \mathcal{RB}(0, \sigma^2),$$

em que ε_n é uma sequência de variáveis aleatórias não correlacionados com média zero e variância constante $\sigma^2 < \infty$. Além disso, ε_n deve ser não correlacionado com X_m para cada $m < n$.

As cadeias desse processo são recorrentes.

Ver arquivo exemplo_44.r

Com o objetivo de gerar da distribuição $f(\boldsymbol{x})$, estamos interessados em cadeias que explorem o espaço de estados $\mathcal{X}$, dessa forma, estamos interessados em cadeias recorrentes.

Se além de recorrente a cadeia tiver **recorrência de Harris**, então

$$\Pr(X_n \in A | x_0) = 1.$$

Essa propriedade garante que começando a cadeia de qualquer ponto inicial a distribuição estacionária é a mesma.

Distribuição invariante

Dizemos que existe a **distribuição invariante** f se a distribuição marginal de X_n não depender de n.

Essa propriedade é necessária para que exista f tal que $X_{n+1} \sim f$ se $X_n \sim f$.

Se f é uma medida de probabilidade, dizemos que a distribuição invariante é estacionária.

Exemplo

Considere o modelo autoregressivo de ordem 1, AR(1), dado por

$$X_n = \theta X_{n-1} + \varepsilon_n, \quad |\theta| < 1 \quad \text{e} \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2).$$

Nesse caso, o núcleo de transição é $\mathcal{N}(\theta x_{n-1}, \sigma^2)$ e essa distribuição é estacionária somente se $|\theta| < 1$.

No caso, $|\theta| < 1$ a distribuição $\mathcal{N}(0, \sigma^2/(1-\theta^2))$ é a única distribuição estacionária do AR(1).

Cadeia reversível

Uma cadeia de Markov é dita **reversível** se a distribuição de $(X_{n+1}|X_{n+2} = x)$ é a mesma de $(X_{n+1}|X_n = x)$.

Essa propriedade está ligada a existência de uma distribuição estacionária f.

Condição balanceada detalhada

Uma cadeia de Markov satisfaz a **condição balanceada detalhada** se existe f tal que

$$K(x|y)f(y) = K(y|x)f(x), \quad \text{para todo} \quad (x, y).$$

Essa propriedade expressa um equilíbrio no movimento $x \to y$ e $y \to x$. A probabilidade de ir de x para y é a mesma de ir de y para x.

Considere uma cadeia de Markov com função de transição K e que satisfaz a condição balanceada detalhada, então

1. a densidade f é a densidade invariante da cadeia; e
2. a cadeia é reversível.

Cadeia ergótica

Para uma cadeia de Markov podemos perguntar qual seria o comportamento para $n \to \infty$.

Isto é, para onde a cadeia está convergindo?

Um candidato natural é a distribuição invariante f.

Para cadeias recorrentes de Harris, com distribuição invariante f, uma partícula α é **ergótica** se

$$\lim_{n \to \infty} |K^n(\alpha|\alpha) - f(\alpha)| = 0.$$

Acoplamento

Considere duas cadeias (X_n) e (Y_n) associadas com o núcleo K. Um evento de acoplamento ocorre quando as duas cadeias se encontram em α, isto é, o primeiro n_0 tal que $X_{n_0} \in \alpha$ e $Y_{n_0} \in \alpha$. Após esse instante (X_n) e (Y_n) são idênticas.

Se mostramos que o tempo para o acoplamento é finito, a ergodicidade da cadeia é satisfeita.

Teorema da ergoticidade: uma cadeia de Markov que é aperiódica, irredutível e recorrente positiva tem uma distribuição limite.

Introdução a Monte Carlo via cadeias de Markov (MCMC)

Objetivo: construir uma cadeia de Markov que tenha como distribuição estacionária $f(\boldsymbol{x})$.

Chamamos método de Monte Carlo via cadeias de Markov (MCMC) o algoritmo para simular de f que usa uma cadeia de Markov (X_n) cuja distribuição estacionária é f.

Ajuda na inferência bayesiana para modelos altamente estruturados e complexos. Neste caso, a distribuição estacionária seria $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Usando Monte Carlo

Suponha que temos uma amostra aleatória $(\theta_1, \ldots, \theta_M)$ gerados da posteriori $\pi(\theta|\boldsymbol{y})$.

Podemos usar Monte Carlo para estudar as propriedades dessa distribuição.

Por exemplo, podemos calcular

$$\mathrm{E}(\theta|\boldsymbol{y}) = \int \theta\pi(\theta|\boldsymbol{y})d\theta \simeq \frac{1}{M}\sum_{i=1}^{M}\theta_i.$$

Intervalo de credibilidade de 0,95 (95%) $\int_{q(0,025)}^{q(0,975)} \pi(\theta|\boldsymbol{y})d\theta$

$$0,025 = \frac{1}{M}\sum_{i=1}^{M} I(\theta_i \leqslant q(0,025)) \quad \text{e} \quad 0,975 = \frac{1}{M}\sum_{i=1}^{M} I(\theta_i \leqslant q(0,975)).$$

Algoritmo de Metropolis-Hastings

Considere uma densidade condicional $q(y|x)$ e uma densidade objetivo $f(x)$.

O algoritmo de Metropolis-Hastings produz uma cadeia de Markov (X_n) através do algoritmo abaixo.

Algoritmo

1. Inicialize x_k;
2. Gere $x^\star$ com distribuição $q(.|x_k)$;
3. Tome

$$x_{k+1} = \begin{cases} x^\star, & \text{com probabilidade } \alpha; \\ x_k, & \text{com probabilidade } 1-\alpha, \end{cases}$$

em que

$$\alpha = \min\left\{1, \frac{f(x^\star)q(x_k|x^\star)}{f(x_k)q(x^\star|x_k)}\right\}.$$

A densidade q é conhecida como densidade proposta e α como a probabilidade de aceitação.

Algoritmo de Metropolis-Hastings

Por que o algoritmo produz uma cadeia cuja distribuição estacionária é $f(x)$?

Veja Chib, S. and Greenberg, E. (1995), “Understanding the Metropolis-Hastings Algorithm”, *The American Statistician*, 49(4), 327-335.

Exemplo

Considere a distribuição

$$f(x) = cx^{-n/2} \exp\left\{-\frac{a}{2x}\right\}.$$

Suponha que se deseja simular uma amostra dessa distribuição para $n = 5$ e $a = 4$ usando o algoritmo de Metropolis.

Considere como densidade proposta a distribuição $\mathcal{U}(0, 100)$.

A probabilidade de aceitação do algoritmo é

$$\alpha = \min\left\{1, \frac{f(x^\star)}{f(x_k)}\right\}.$$

Suponha que a cadeia é inicializada em $x = 1$ e propõe-se o valor $x = 39,82$ (gerado da $\mathcal{U}(0, 100)$).

Então, $\alpha = 0,0007$, isto é, a probabilidade de aceitar mover a cadeia de 1 para 39,82 é muito pequena.

Ver arquivo exemplo_45.r

Algoritmo de Metropolis-Hastings

Considere uma densidade condicional $q(y|x)$ e uma densidade objetivo $f(x)$.

O algoritmo de Metropolis-Hastings produz uma cadeia de Markov (X_n) através do algoritmo abaixo.

Algoritmo

1. Inicialize x_k;
2. Gere $x^\star$ com distribuição $q(.|x_k)$;
3. Tome

$$x_{k+1} = \begin{cases} x^\star, & \text{com probabilidade } \alpha; \\ x_k, & \text{com probabilidade } 1-\alpha, \end{cases}$$

em que

$$\alpha = \min\left\{1, \frac{f(x^\star)q(x_k|x^\star)}{f(x_k)q(x^\star|x_k)}\right\}.$$

A densidade $q(.|.)$ é conhecida como densidade proposta e α como a probabilidade de aceitação.

Escolha da proposta

A escolha da proposta de transição $q(.|.)$ é crucial para a implementação do algoritmo de Metropolis-Hastings.

As propostas mais usadas são:

- **passeio aleatório**; e
- **proposta independente**.

Proposta passeio aleatório

Uma maneira prática de construir um algoritmo de Metropolis-Hastings é usar o valor de x em que a cadeia está para propor um movimento da cadeia.

Isto é, considerar uma exploração da vizinhança em torno do valor atual da cadeia.

Note que a densidade proposta pode depender do valor atual da cadeia, $q(.|x_k)$. Por exemplo, podemos considerar

$$X^\star = X_k + \varepsilon,$$

em que ε uma perturbação aleatória simétrica em torno de 0.

Dessa forma, $q(x^\star|x_k) = g(|x^\star - x_k|)$, isto é, $g(.)$ é simétrica em torno de x_k.

A probabilidade de aceitação é dada por

$$\alpha = \min\left\{1, \frac{f(x^\star)}{f(x_k)}\right\} \quad \text{pois} \quad q(x^\star|x_k) = q(x_k|x^\star).$$

Note ainda que para $f(x^\star) > f(x_k)$ a proposta é automaticamente aceita.

Essa é uma proposta muito usada na prática. Em geral, usamos a distribuição normal como proposta.

Note que, neste caso, precisamos ajustar a variância da distribuição.

- Se a variância é pequena, aceitamos mais movimentos, porém os movimentos são lentos no suporte de X.
- Por outro lado, se a variância é grande, aceitamos menos movimentos, porém visitamos mais rapidamente o suporte de X.

Exemplo

Suponha que queremos uma amostra de uma mistura de normais bivariadas,

$$f(\boldsymbol{x}) = 0,7\phi(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0,3\phi(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad \text{com}$$

$\boldsymbol{\mu}_1 = (4; 5)^\top$, $\boldsymbol{\mu}_2 = (0,7; 3,5)^\top$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0,7 \\ 0,7 & 1 \end{pmatrix}$ e

$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0,7 \\ -0,7 & 1 \end{pmatrix}$.

Suponha que não sabemos fazer isto diretamente (Monte Carlo simples).

Considere um algoritmo de Metropolis-Hastings com proposta de passeio aleatório,

$$\boldsymbol{x}^\star \sim \mathcal{N}_p(\boldsymbol{x}^{(i-1)}, \boldsymbol{\Psi}),$$

para alguma matriz de covariâncias $\boldsymbol{\Psi}$ adequada.

Ver arquivo exemplo_46.r

Cadeias independentes

Nesse caso, a proposta não depende do valor da cadeia no passo anterior,

$$q(x|x_k) = g(x).$$

Note que ainda temos uma cadeia de Markov porque a probabilidade de aceitação em cada passo depende do passo anterior,

$$\alpha = \min\left\{1, \frac{f(x^\star)g(x_k)}{f(x_k)g(x^\star)}\right\}.$$

Propostas que considerem a forma da distribuição de interesse terão vantagens nesse caso. Por exemplo, aproximações pela normal.

Exemplo (continuação)

Agora, utilizaremos uma proposta independente $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ com $\boldsymbol{\mu} = (3, 5)^\top$ e $\boldsymbol{\Sigma} = 2\boldsymbol{I}_2$.

Ver arquivo exemplo_47.r

Exemplo

O total de n animais são divididos em 3 categorias $\boldsymbol{y} = (y_1, y_2, y_3)^\top$ com probabilidades $(0,25(2+\theta);\ 0,5(1-\theta),\ 0,25\theta)$.

Temos a função de verossimilhança

$$f(\boldsymbol{y}|\theta) \propto (2+\theta)^{y_1}(1-\theta)^{y_2}\theta^{y_3}, \quad \theta \in (0,1).$$

Considere a priori $\theta \sim \mathcal{U}(0,1)$. Observou-se $\boldsymbol{y} = (125,\ 38,\ 34)^\top$. Como encontrar a distribuição a posteriori de θ?

- Consideraremos Metropolis-Hastings com proposta independente $\mathcal{U}(0,1)$.

  Ver arquivo exemplo_48.r

- Sabemos que $\theta \in (0,1)$. Como usar uma proposta passeio aleatório nesse caso?

  Ver arquivo exemplo_49.r

  Faremos $\theta = \exp\{\xi\}/(1+\exp\{\xi\})$ e $\xi = \log(\theta/(1-\theta))$ com $\xi^\star \sim \mathcal{N}(\xi^{(i-1)}, V)$. Note que $\dfrac{d\xi}{d\theta} = \dfrac{1}{\theta(1-\theta)}$ e $\dfrac{d\theta}{d\xi} = \dfrac{\exp\{\xi\}}{(1+\exp\{\xi\})^2}$.

Exemplo

Acredita-se que a temperatura afeta a probabilidade de falha de um componente de foguetes. Por exemplo, acredita-se que em um acidente de 1986 a falha ocorreu devido a uma temperatura muito abaixo de 0 graus.

Considere o modelo de regressão logística para a probabilidade de falha dessa componente,

$$(y_i|\pi_i) \sim \text{Ber}(\pi_i) \quad \text{com} \quad \pi_i = \frac{\exp\{\beta_1 + \beta_2 x_i\}}{1 + \exp\{\beta_1 + \beta_2 x_i\}}.$$

Considere distribuição a priori $\mathcal{N}(0, 100)$ para ambos os parâmetros, β_1 e β_2, independentemente.

Como usar o algoritmo de Metropolis-Hastings com proposta de passeio aleatório para obter uma amostra da posteriori de (β_1, β_2)?

Ver arquivo exemplo_50.r

Algoritmo de Gibbs

É um esquema de Monte Carlo via cadeias de Markov em que a proposta de transição é a **distribuição condicional completa**.

No algoritmo de Gibbs as propostas são sempre aceitas.

O teorema a seguir nos diz que ter as distribuições condicionais é suficiente para recuperarmos a distribuição conjunta.

Dessa forma, ao invés de gerar valores de uma distribuição conjunta, podemos gerar valores das condicionais.

Teorema

A distribuição conjunta associada com as densidades condicionais $f_1(y|x)$ e $f_2(x|y)$ é

$$f(x,y) = \frac{f_1(y|x)}{\int \frac{f_1(y|x)}{f_2(x|y)}\,dy}.$$

Núcleo de transição

Sob certas condições podemos mostrar que a distribuição estacionária do núcleo de transição abaixo é a densidade conjunta de $f(x, y)$,

$$K((x, y), (x', y')) = f_1(y'|x')f_2(x'|y).$$

Estamos interessados na distribuição $f(\boldsymbol{x})$ com $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^\top$.

Cada componente de $\boldsymbol{x}_i$ pode ser um escalar, um vetor ou uma matriz.

Suponha que as **distribuições condicionais completas** estejam disponíveis:

$$f(\boldsymbol{x}_i|\boldsymbol{x}_{-i}), \quad i = 1, \ldots, p,$$

em que $\boldsymbol{x}_{-i} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_p)^\top$.

A ideia é que gerar de $f(\boldsymbol{x}_i|\boldsymbol{x}_{-i})$ é menos custoso (computacionalmente) do que gerar da conjunta $f(\boldsymbol{x})$.

Algoritmo de Gibbs

Algoritmo

1. Inicialize $\boldsymbol{x}^{(0)} = (\boldsymbol{x}_1^{(0)}, \ldots, \boldsymbol{x}_p^{(0)})^\top$.
2. Amostre de cada condicional completa iterativamente:
   - 2.1 $\boldsymbol{x}_1^{(j)}$ com densidade $f(\boldsymbol{x}_1 | \boldsymbol{x}_2^{(j-1)}, \ldots, \boldsymbol{x}_p^{(j-1)})$;
   - 2.2 $\boldsymbol{x}_2^{(j)}$ com densidade $f(\boldsymbol{x}_2 | \boldsymbol{x}_1^{(j)}, \boldsymbol{x}_3^{(j-1)}, \ldots, \boldsymbol{x}_p^{(j-1)})$;
   - 2.3 $\boldsymbol{x}_3^{(j)}$ com densidade $f(\boldsymbol{x}_3 | \boldsymbol{x}_1^{(j)}, \boldsymbol{x}_2^{(j)}, \boldsymbol{x}_4^{(j-1)}, \ldots, \boldsymbol{x}_p^{(j-1)})$; e
   - 2.4 assim por diante até $\boldsymbol{x}_p^{(j)}$ com densidade $f(\boldsymbol{x}_p | \boldsymbol{x}_1^{(j)}, \ldots, \boldsymbol{x}_{p-1}^{(j)})$.

Atenção: cuidado com os índices!

O algoritmo de Gibbs é equivalente a composição de p algoritmos de Metropolis-Hastings com probabilidade de aceitação igual a 1.

Exemplo

Seja $\boldsymbol{X}$ um vetor aleatório d-dimensional com distribuição normal (multivariada) com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$. Então, a função de densidade de probabilidade é dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad \boldsymbol{x}\in\mathbb{R}^p,$$

em que $\boldsymbol{\mu}\in\mathbb{R}^p$ e $\boldsymbol{\Sigma}$ é uma matriz positiva definida. Sejam

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \quad \text{dimensões} \quad \begin{bmatrix} q\times 1\\ (p-q)\times 1\end{bmatrix}, \quad \text{e}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}, \quad \text{dimensões} \quad \begin{bmatrix} q\times q & q\times(p-q)\\ (p-q)\times q & (p-q)\times(p-1)\end{bmatrix}.$$

Então, $(\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{\xi})\sim\mathcal{N}_q(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})$, em que

$$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu}_2) \quad \text{e} \quad \widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Ver arquivo exemplo_51.r

Exemplo

Considere o modelo

$$(y_i|\mu,\phi) \sim \mathcal{N}(\mu,\phi^{-1}), \quad i = 1,\ldots,n,$$

para (μ,ϕ) desconhecidos.

Se considerarmos a distribuição a priori

$$\mu \sim \mathcal{N}(0,\tau_0^2) \quad \text{e} \quad \phi \sim \mathcal{G}\left(\frac{a_0}{2},\frac{b_0}{2}\right),$$

então as distribuições condicionais completas são

$$\begin{aligned}(\phi|\boldsymbol{y},\mu) &\sim \mathcal{G}\left(\frac{n+a_0}{2},\frac{b_0+\sum_{i=1}^n (y_i-\mu)^2}{2}\right)\\ (\mu|\boldsymbol{y},\phi) &\sim \mathcal{N}(\xi,\delta^2),\end{aligned}$$

com $\delta^2 = (n\phi+\tau_0^{-2})^{-1}$ e $\xi = n\overline{y}\phi\delta^2$.

Ver arquivo exemplo_52.r

Exemplo

Seja $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ uma amostra aleatória da Poisson. Acredita-se ter um ponto de mudança m - que é desconhecido - entre as n observações.

Seja $(y_1, \ldots, y_m)$ tal que $y_i \sim \text{Poi}(\lambda)$.

Seja $(y_{m+1}, \ldots, y_n)$ tal que $y_i \sim \text{Poi}(\phi)$.

Os parâmetros do modelo são (λ, ϕ, m).

Objetivo: dada a amostra, obter a distribuição a posteriori para os parâmetros de interesse, isto é,

$$f(\lambda, \phi, m|\boldsymbol{y}).$$

Assumimos que (λ, ϕ, m) são independentes a priori com

$$\begin{aligned} \lambda &\sim \mathcal{G}(a, b), \\ \phi &\sim \mathcal{G}(c, d) \quad \text{e} \\ f(m) &= \frac{1}{n-1} \quad \text{para} \quad m = 1, \ldots, n-1. \end{aligned}$$

A distribuição a posteriori é

$$\begin{aligned} f(\lambda, \phi, m|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\lambda, \phi, m) f(\lambda, \phi, m) \\ &= \left[\prod_{i=1}^{m} f(y_i|\lambda)\right] \left[\prod_{i=m+1}^{n} f(y_i|\phi)\right] f(\lambda|a, b) f(\phi|c, d) f(m) \\ &\propto \lambda^{a+s_1(m)-1} \phi^{c+s_2(m)-1} \exp\{-(b+m)\lambda - (d+(n-m))\phi\}, \end{aligned}$$

em que $s_1(m) = \sum_{i=1}^{m} y_i$ e $s_2(m) = \sum_{i=m+1}^{n} y_i$.

Neste exemplo, a posteriori conjunta é complicada, porém as distribuições condicionais completas são relativamente simples:

$$\begin{aligned} (\lambda|\boldsymbol{y}, \phi, m) &\sim \mathcal{G}(a + s_1(m), b + m) \\ (\phi|\boldsymbol{y}, \lambda, m) &\sim \mathcal{G}(c + s_2(m), d + n - m) \\ f(m|\boldsymbol{y}, \lambda, \phi) &= \frac{g(m)}{\sum_{j=1}^{n-1} g(j)}, \quad \text{para} \quad m = 1, \ldots, n-1, \end{aligned}$$

em que $g(m) = \lambda^{s_1(m)} \phi^{s_2(m)} \exp\{m(\phi - \lambda)\}$.

Ver arquivo exemplo_53.r

Implementação: obtendo a amostra

Uma maneira de obter uma amostra de tamanho N da distribuição a posteriori é considerar N cadeias independentes começando de pontos diferentes.

Se todas as cadeias têm indícios de convergência a partir da iteração M, pegamos o M-ésimo valor de cada cadeia i para compor a amostra.

Neste caso, precisaremos de $N \times M$ gerações das cadeias conjuntamente.

Outra opção é considerar uma única cadeia e obter N pontos após indícios de convergência (na M-ésima iteração).

Neste caso, precisaremos de $N + M$ gerações da cadeia. Neste caso é mais eficiente do que $N \times M$ gerações.

Implementação: autocorrelação

No caso de uma única cadeia tomamos N valores após os indícios de convergência, mas os valores gerados não são independentes.

Para obter valores que sejam menos autocorrelacionados, podemos tomar uma observação a cada k valores gerados da cadeia.

Dessa forma, precisaremos de $M + kN$ valores gerados da cadeia. O que ainda é mais eficiente do que $N \times M$ gerações usando N cadeias independentes.

Uma outra opção é usar um número pequeno (R) de cadeias independentes e após os indícios de convergência tomar um a cada k valores gerados.

Precisaremos gerar $R(M + kN/R) = RM + kN$ valores das cadeias.

Em cada exemplo a seguir, verificar **a função de autocorrelação (ACF)**.

Ver arquivo exemplo_54.r

Ver arquivo exemplo_55.r

Implementação: cadeias paralelas?

Usar *R* cadeias paralelas é de forma geral computacionalmente muito ineficiente.

Muitas vezes será necessário usar algumas cadeias em paralelo.

Para ser prudente, podemos usar poucas cadeias paralelas. Se elas se misturam em torno de um valor, então é “seguro” usar uma única cadeia.

Contudo, se modas locais são visitadas em alguma(s) da(s) cadeia(s) significa que precisamos de longas cadeias e o uso de algumas cadeias em paralelo é recomendado.

Ver arquivo exemplo_55.r

Implementação: amostragem em blocos

Para obter amostras da distribuição de um vetor $\boldsymbol{x}$ podemos usar um único bloco ou arranjar os parâmetros em blocos.

Por exemplo, se amostramos x_i individualmente temos blocos unidimensionais.

Blocos maiores permitem movimentos em direções mais variadas.

Se os parâmetros são muito correlacionados, devemos amostrá-los em um único bloco.

Implementação: convergência

Vimos que um valor da distribuição estacionária é obtido para o número de iterações indo para infinito.

Na prática consideramos um número grande de iterações.

Quando devemos considerar que temos um número grande de iterações?

Chamaremos de aquecimento (*burn-in* ou *warm-up*) o período até obtermos indícios de convergência da cadeia.

Nem sempre é fácil verificar se a "convergência" foi obtida.

Implementação: quando parar?

Algumas questões importantes que devem ser levadas em consideração:

- Estacionariedade: estamos explorando todo o domínio da distribuição de interesse?
- Estimação: somos capazes de estimar os parâmetros de interesse dada uma certa precisão?

Implementação: convergência

Podemos verificar convergência de duas formas:

1. estudando características teóricas da cadeia e verificando se ela está próxima da distribuição objetivo.
2. estudando a cadeia sob o ponto de vista estatístico. Isto é, estudando as características da cadeia observada.

Pelo Item (2), podemos obter métodos práticos de verificar “convergência” . Porém esta não é garantida, pois analisamos apenas a cadeia observada e não as características teóricas.

Implementação: verificação informal

Indícios de convergência podem ser verificados usando várias cadeias em paralelo.

Média, desvio padrão, etc. devem se manter os mesmos em todas as cadeias para que isto se caracteriza como os indícios de convergência.

Se as sequências dos valores gerados para cada variável nas várias cadeias se mantêm em torno de valores constantes (níveis), então diremos que a cadeia tem indícios de convergência (a cadeia convergiu).

Implementação: tamanho do aquecimento

Raftery e Lewis (1992) propuseram um método para estabelecer o tamanho da cadeia (período de aquecimento) até os indícios de convergência.

O método sugere o valor de M (aquecimento), a defasagem k (tomamos uma observação a cada k iterações) e N (tamanho da amostra para obter certa precisão na estimação de Monte Carlo).

O método é baseado em exigir que

$$\Pr(|\hat{q} - q| \leqslant r) = s$$

tal que $q = \Pr(\psi(x) \leqslant u)$.

Ver arquivo exemplo_56.r

Ver arquivo exemplo_57.r

Métodos formais de convergência

Considere uma função $\phi = t(\boldsymbol{x})$.

A trajetória $\phi^{(1)}, \ldots, \phi^{(K)}$ forma uma série temporal.

Considere uma série com tamanho $K = M + N$ e médias

$$\begin{aligned} \overline{\phi}_a &= \frac{1}{n_a} \sum_{j=M+1}^{M+n_a} \phi^{(j)} \\ \overline{\phi}_b &= \frac{1}{n_b} \sum_{j=M+N-n_b+1}^{M+N} \phi^{(j)}. \end{aligned}$$

Temos $\overline{\phi}_a$ e $\overline{\phi}_b$ como as médias do começo e do fim da série após convergência (em M) e $n_a + n_b < N$.

Métodos formais de convergência

Usando o teorema central do limite, temos que

$$Z_g = \frac{\overline{\phi}_a - \overline{\phi}_b}{\sqrt{\widehat{V}(\overline{\phi}_a) + \widehat{V}(\overline{\phi}_b)}} \approx \mathcal{N}(0,1).$$

Então, o valor observado de Z_g não deve ser grande no caso de convergência da série.

Valores grandes indicam que ainda não temos indícios de convergência da cadeia.

Uma sugestão considera $t(\boldsymbol{x}) = -2\log(f(\boldsymbol{x}))$, $n_a = 0,5n$ e $n_b = 0,1n$.

Ver arquivo exemplo_56.r

Ver arquivo exemplo_57.r

Teorema fundamental da simulação

Considere a ideia que esta por trás de vários métodos de simulação tais como rejeição e amostragem da fatia (*Slice Sampler*). Podemos reescrever a densidade objetivo $f(x)$ por

$$f(x) = \int_0^{f(x)} 1 du,$$

isto é, $f(x)$ é a densidade marginal da conjunta de (X, U) em que

$$(X, U) \sim \mathcal{U}\{(x, u) : \ 0 < u < f(x)\}.$$

Dessa forma

$$f(x) = \int I_{(0,f(x))}(u) du = \int_0^{f(x)} 1 du.$$

A variável U é chamada de variável auxiliar, criada artificialmente para facilitar a geração de variáveis aleatórias com distribuição $f(x)$.

Para gerar X com função de densidade de probabilidade $f(x)$ basta gerar da conjunta de (X, U).

Para isto usamos uma cadeia de Markov, via algoritmo de Gibbs, que gera valores de $(X|U)$ e de $(U|X)$ tais que

$$\begin{aligned}(X|U=u) &\sim \mathcal{U}(x:\ u \leqslant f(x)) \\ (U|X=x) &\sim \mathcal{U}(u:\ u \leqslant f(x)).\end{aligned}$$

Os valores gerados são obtidos da distribuição conjunta de (X, U).

Teorema

Gerar X com função de densidade de probabilidade $f(x)$ é equivalente a gerar $(X, U) \sim \mathcal{U}\{(x, u):\ 0 < u < f(x)\}$.

Movemos a cadeia fazendo um movimento no eixo u. Depois, fazemos um movimento no eixo x.

Após gerar n pares (x_i, u_i), basta considerar somente os valores de $x_1, \ldots, x_n$ como amostra da marginal de X.

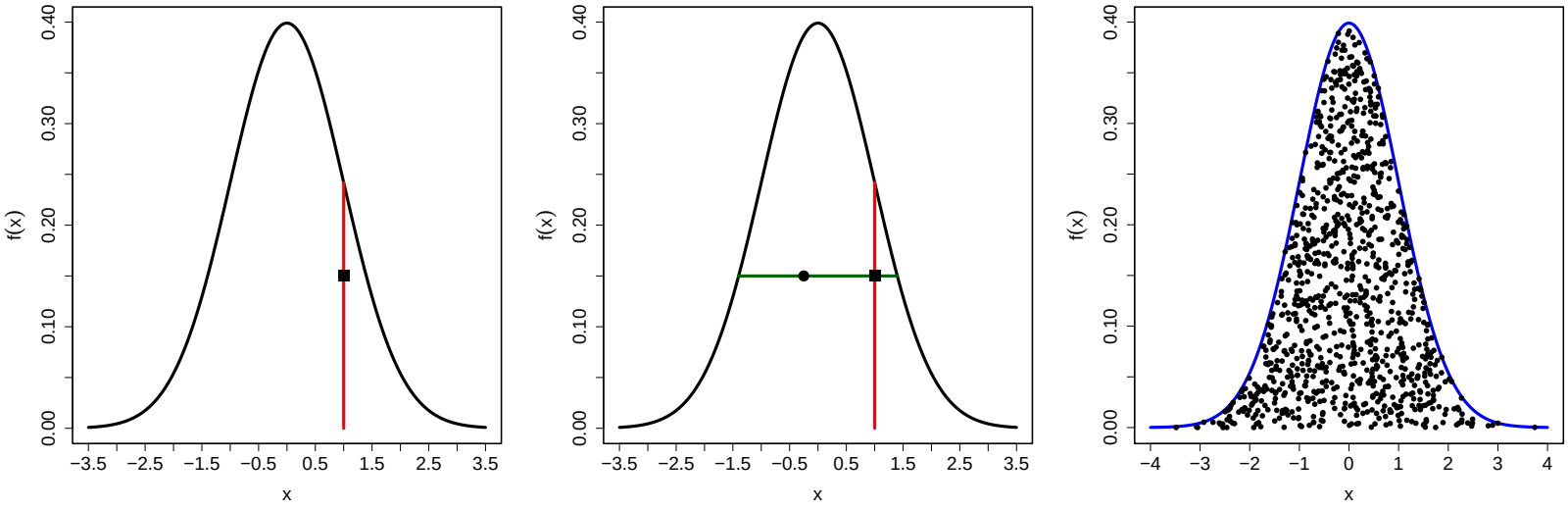


Figura 12: Comportamento da amostragem da fatia.

Esse método gera valores somente da distribuição uniforme.

Nenhum valor é rejeitado.

Ideia: gerar valores sob a curva baseados em uniformes.

Suponha uma distribuição $f(x)$ unimodal.

Algoritmo

1. Fazer $i = 0$ e inicializar com x_0;
2. Fazer $i = i + 1$ e gerar y_i de $\mathcal{U}(0, f(x_{i-1}))$;
3. Gerar x_i de $\mathcal{U}(c_{1i}, c_{2i})$ em que $y_i = f(c_{1i}) = f(c_{2i})$;
4. Repetir N vezes os Passos 2 e 3.

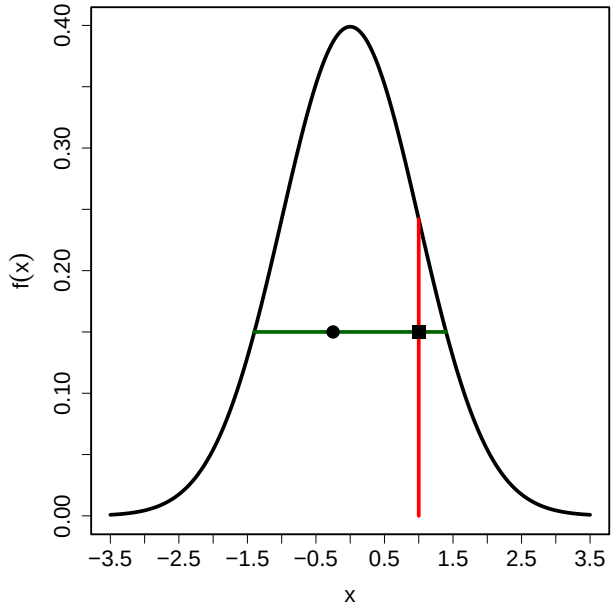


Figura 13: Comportamento da amostragem da fatia.

Ver arquivo exemplo_58.r

Exemplo

Desejamos obter uma amostra da $\mathcal{N}(0,1)$ truncada no $(0,1)$ via amostragem da fatia. A função de densidade de probabilidade é dada por

$$f(x) = \frac{\phi(x)}{\Phi(1) - \Phi(0)}, \quad x \in (0,1).$$

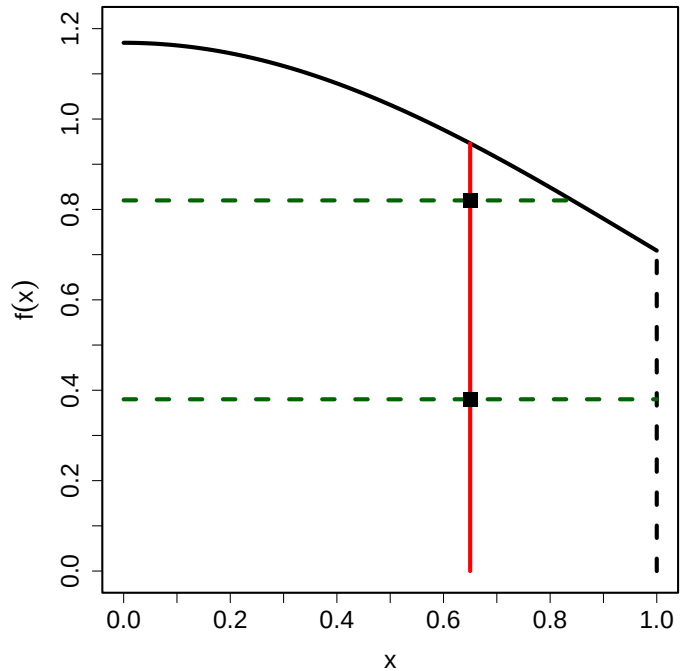


Figura 14: Comportamento da amostragem da fatia para normal truncada.

Dado y, precisamos calcular κ tal que $f(\kappa) = y$:

$$\kappa = \sqrt{-2\left[\log(y) + \frac{1}{2}\log(2\pi) + \log(\Phi(1) - \Phi(0))\right]}.$$

Note que $0 \leqslant \kappa \leqslant 1$.

Algoritmo

1. Fazer $i = 0$ e inicializar com x_0;
2. Fazer $i = i + 1$ e gerar y_i de $\mathcal{U}(0, f(x_{i-1}))$;
3. Gerar x_i de $\mathcal{U}(0, \min(\kappa_i, 1))$.
4. Repetir N vezes os Passos 2 e 3.

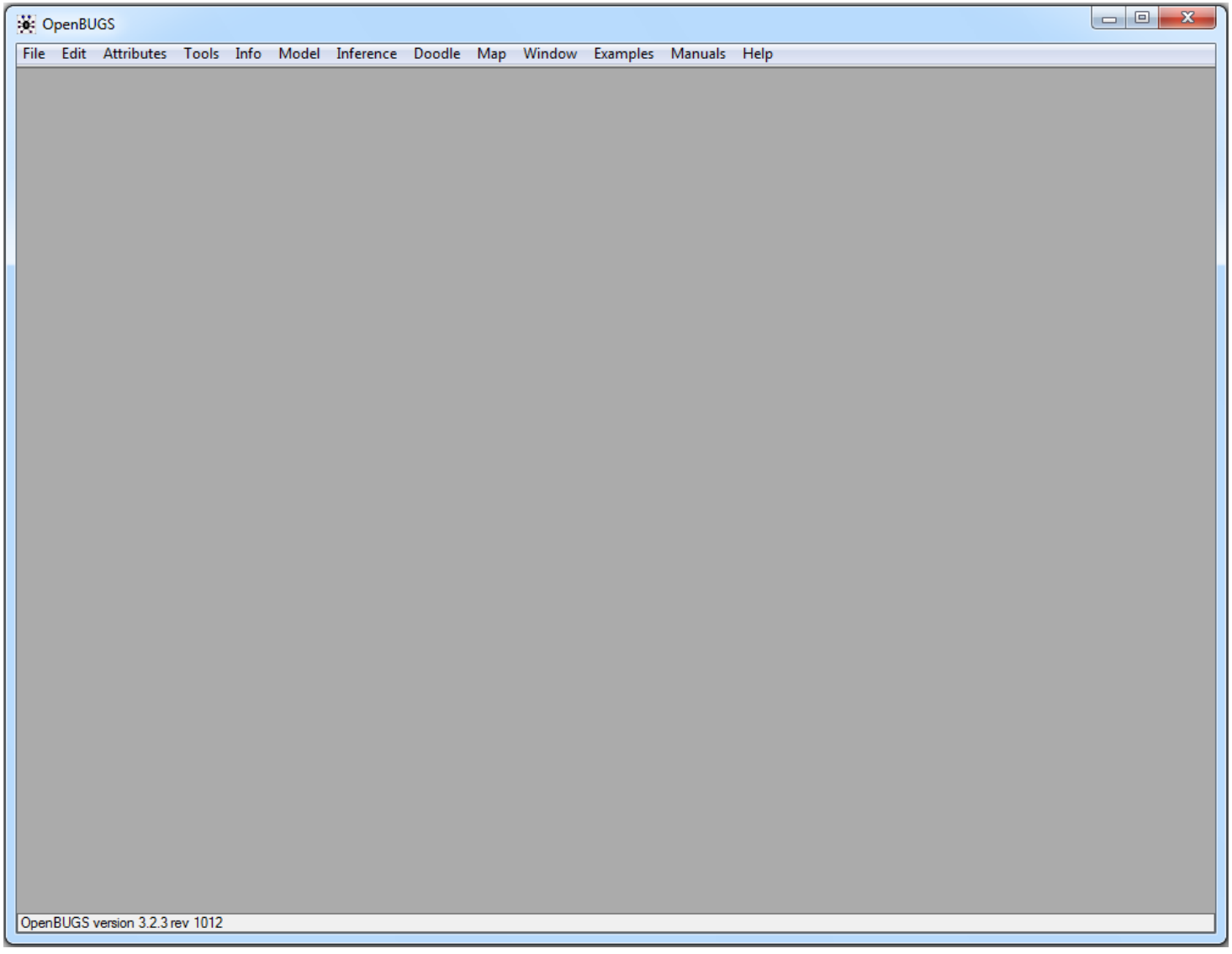
Ver arquivo exemplo_59.r

O OpenBUGS

- Bastante amigável desde o ponto de vista do usuário;
- **OpenBUGS** disponível no sítio the-bugs-project-openbugs;
- É uma ferramenta ideal para a modelagem;
- Permite mudar a especificação da distribuição a priori de forma simples;
- Em algumas aplicações a convergência pode ser lenta.

Antes, utilizamos o **WinBUGS** disponível the-bugs-project-winbugs;

O entorno do OpenBUGS



Exemplo

Considere o modelo de regressão linear múltiplo dado por:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \quad t = 1, 2, \ldots, n.$$

A distribuição a posteriori é dada por:

$$f(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} f(\boldsymbol{\beta}, \sigma^2),$$

em que $f(\boldsymbol{\beta}, \sigma^2)$ é a distribuição a priori, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p})$, $\mathbf{x_i} = (x_{i1}, \ldots, x_{in})^\top$, e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

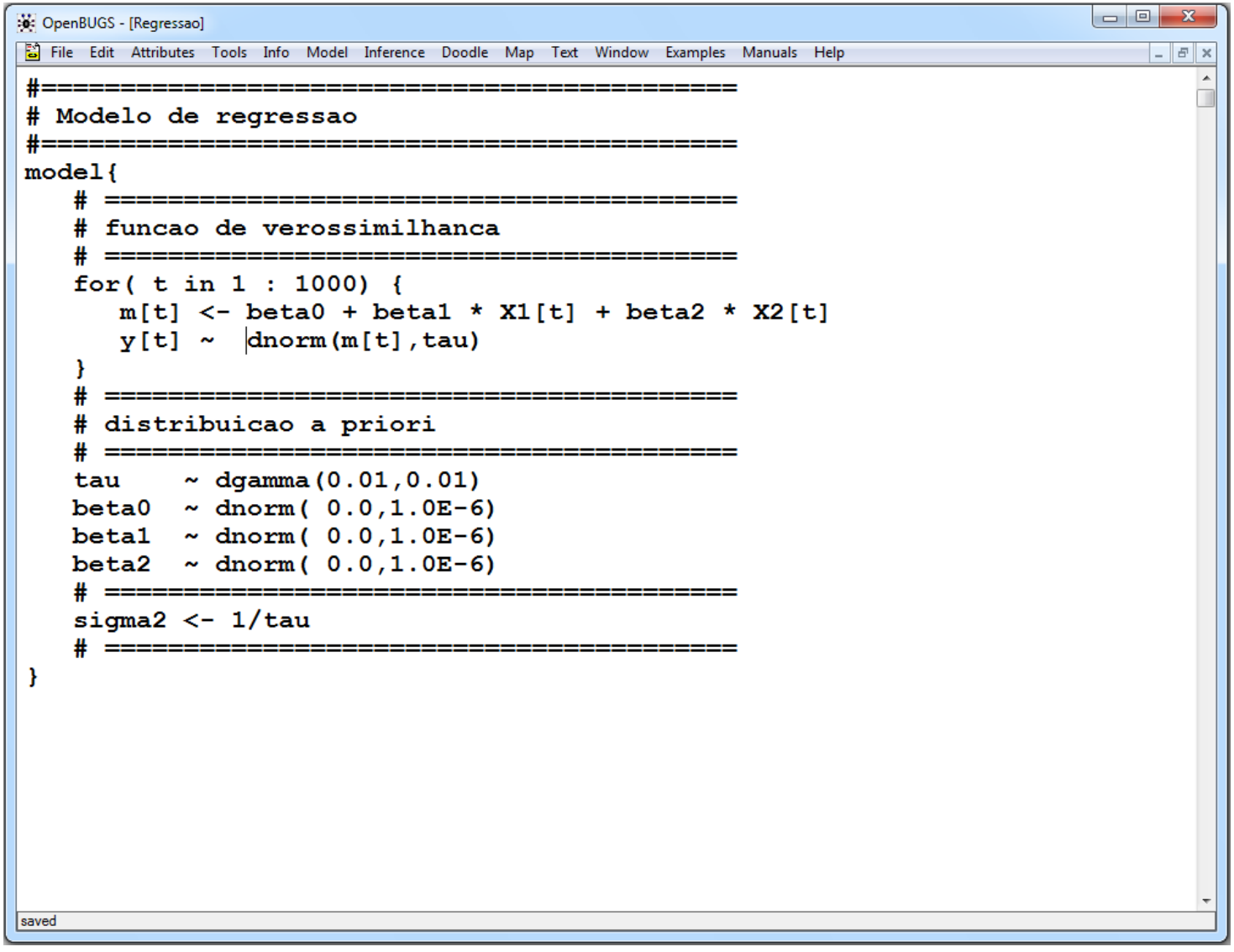
Dados artificiais gerados do seguinte modelo:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad t = 1, 2, \ldots, n,$$

com $\beta_0 = 1,0$,$\beta_1 = 5,0$, $\beta_2 = -8,0$, $\sigma^2 = 0,25$ e $n = 1000$.

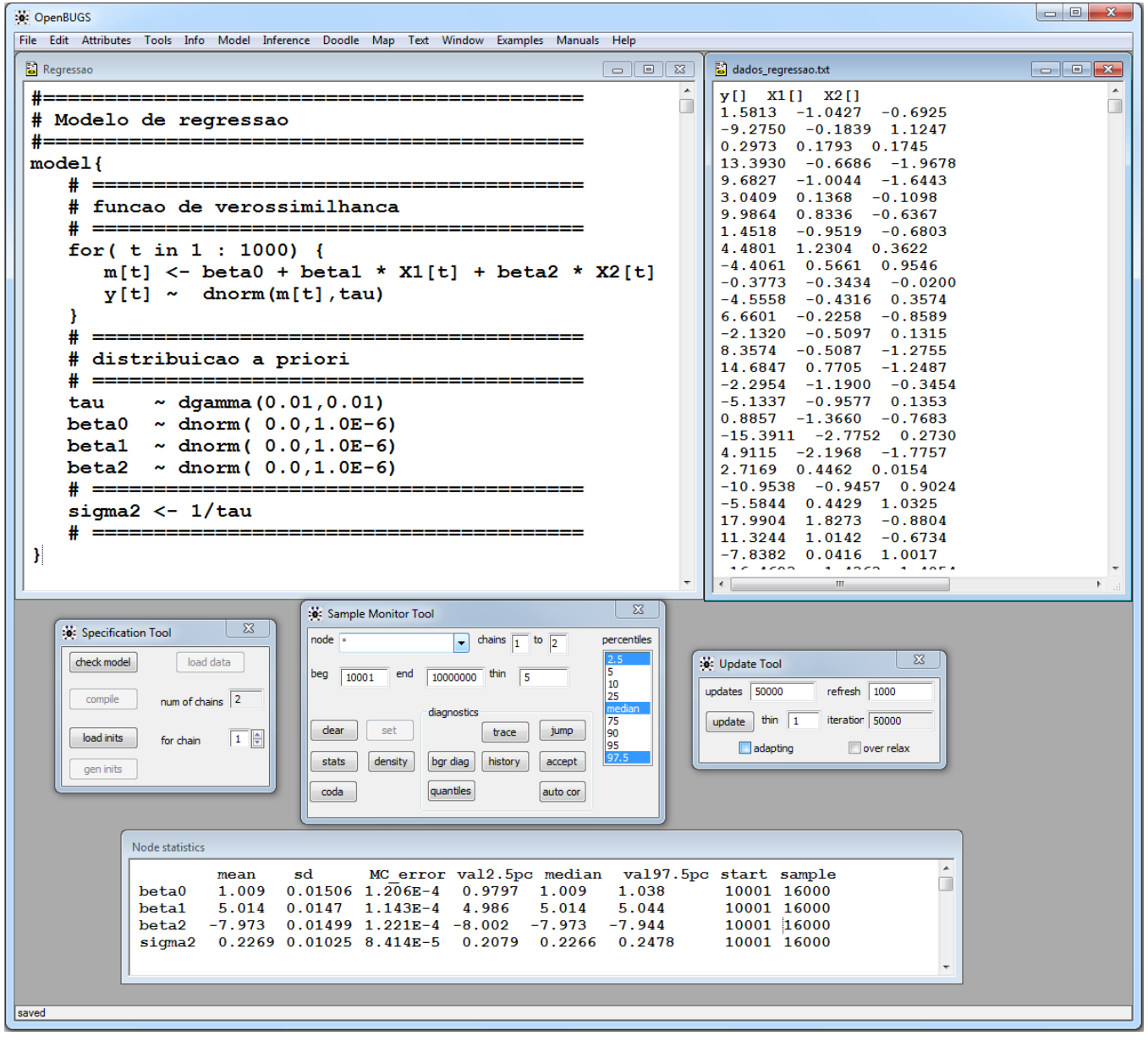
O OpenBUGS trabalha com a precisão $\phi = 1/\sigma^2$.

O modelo em código BUGS

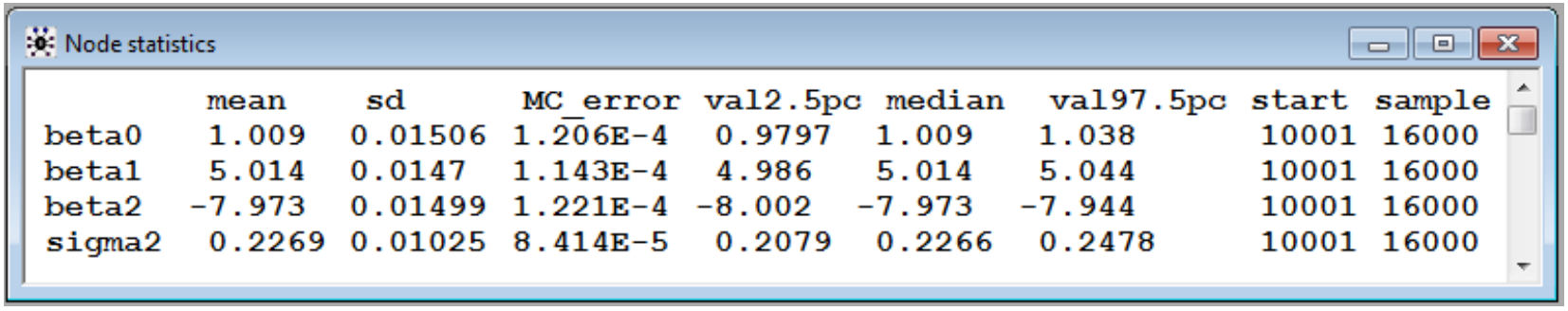


Opções básicas do OpenBUGS

1. ***Model*→ *Specification...***
2. ***check model***
   Verificar por *model is syntactically correct*
3. ***load data***
   Verificar por *data loaded*
4. ***compile***
   Verificar por *model compiled*
5. ***load inits*** ou ***gen inits***
   Verificar por *initial values generated, model initialized*
6. ***Inference*→ *Samples...***
7. ***Model*→ *Update...***


OpenBUGS
File Edit Attributes Tools Info Model Inference Doodle Map Text Window Examples Manuals Help
Regressao
#===
Modelo de regressao
#===
model{
# ===
# funcao de verossimilhanca
# ===
for(t in 1 : 1000) {
m[t] <- beta0 + beta1 * X1[t] + beta2 * X2[t]
y[t] ~ dnorm(m[t],tau)
}
# ===
# distribuicao a priori
# ===
tau ~ dgamma(0.01,0.01)
beta0 ~ dnorm(0.0,1.0E-6)
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
# ===
sigma2 <- 1/tau
# ===
}
dados_regressao.txt
y[] X1[] X2[]
1.5813 -1.0427 -0.6925
-9.2750 -0.1839 1.1247
0.2973 0.1793 0.1745
13.3930 -0.6686 -1.9678
9.6827 -1.0044 -1.6443
3.0409 0.1368 -0.1098
9.9864 0.8336 -0.6367
1.4518 -0.9519 -0.6803
4.4801 1.2304 0.3622
-4.4061 0.5661 0.9546
-0.3773 -0.3434 -0.0200
-4.5558 -0.4316 0.3574
6.6601 -0.2258 -0.8589
-2.1320 -0.5097 0.1315
8.3574 -0.5087 -1.2755
14.6847 0.7705 -1.2487
-2.2954 -1.1900 -0.3454
-5.1337 -0.9577 0.1353
0.8857 -1.3660 -0.7683
-15.3911 -2.7752 0.2730
4.9115 -2.1968 -1.7757
2.7169 0.4462 0.0154
-10.9538 -0.9457 0.9024
-5.5844 0.4429 1.0325
17.9904 1.8273 -0.8804
11.3244 1.0142 -0.6734
-7.8382 0.0416 1.0017
Specification Tool
check model
load data
compile
num of chains 2
load inits
for chain 1
gen inits
Sample Monitor Tool
node *
chains 1 to 2
beg 10001
end 10000000
thin 5
percentiles
2.5
5
10
25
median
75
90
95
97.5
diagnostics
clear
set
trace
jump
stats
density
bgr diag
history
accept
coda
quantiles
auto cor
Update Tool
updates 50000
refresh 1000
update
thin 1
iteration 50000
adapting
over relax
Node statistics
mean sd MC_error val2.5pc median val97.5pc start sample
beta0 1.009 0.01506 1.206E-4 0.9797 1.009 1.038 10001 16000
beta1 5.014 0.0147 1.143E-4 4.986 5.014 5.044 10001 16000
beta2 -7.973 0.01499 1.221E-4 -8.002 -7.973 -7.944 10001 16000
sigma2 0.2269 0.01025 8.414E-5 0.2079 0.2266 0.2478 10001 16000
saved

Resultados

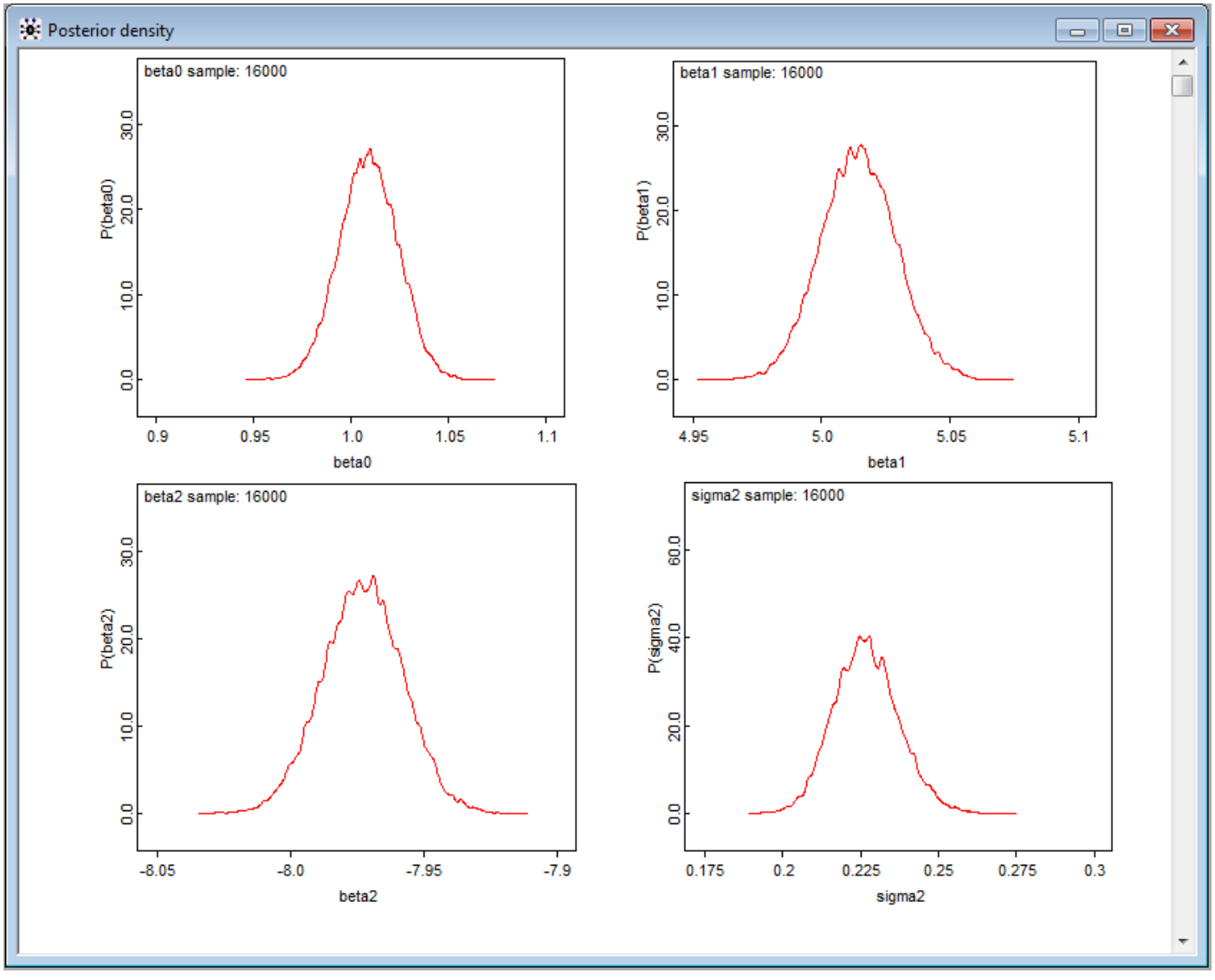

Node statistics

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
beta0	1.009	0.01506	1.206E-4	0.9797	1.009	1.038	10001	16000
beta1	5.014	0.0147	1.143E-4	4.986	5.014	5.044	10001	16000
beta2	-7.973	0.01499	1.221E-4	-8.002	-7.973	-7.944	10001	16000
sigma2	0.2269	0.01025	8.414E-5	0.2079	0.2266	0.2478	10001	16000

Parâmetro	Média	D.P.	2,5%	Mediana	97,5%
β_0	1,009	0,0151	0,980	1,009	1,039
β_1	5,014	0,0147	4,986	5,014	5,043
β_2	-7,973	0,0150	-8,002	-7,973	-7,944
σ^2	0,227	0,0103	0,208	0,227	0,248

Ver arquivos regressao.odc e dados_regressao.txt

Resultados


Posterior density
beta0 sample: 16000
P(beta0)
0.0
10.0
20.0
30.0
0.9
0.95
1.0
1.05
1.1
beta0
beta1 sample: 16000
P(beta1)
0.0
10.0
20.0
30.0
4.95
5.0
5.05
5.1
beta1
beta2 sample: 16000
P(beta2)
0.0
10.0
20.0
30.0
-8.05
-8.0
-7.95
-7.9
beta2
sigma2 sample: 16000
P(sigma2)
0.0
20.0
40.0
60.0
0.175
0.2
0.225
0.25
0.275
0.3
sigma2

Exemplo

Considere o seguinte modelo de regressão logística:

$$\begin{aligned} f(y_t = \omega|\boldsymbol{\beta}) &= \pi_t^{\omega}(1-\pi_t)^{1-\omega}, \quad \omega = 0, 1 \\ \text{logit}(\pi_t) &= \log\left(\frac{\pi_t}{1-\pi_t}\right) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt}, \end{aligned}$$

para $t = 1, 2, \ldots, n$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\top}$.

A distribuição a posteriori é dada por:

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\left\{\sum_{t=1}^{n}\left[y_t \mathbf{x}_t^{\top}\boldsymbol{\beta} - \log(1 + \exp\{\mathbf{x}_t^{\top}\boldsymbol{\beta}\})\right]\right\} f(\boldsymbol{\beta}),$$

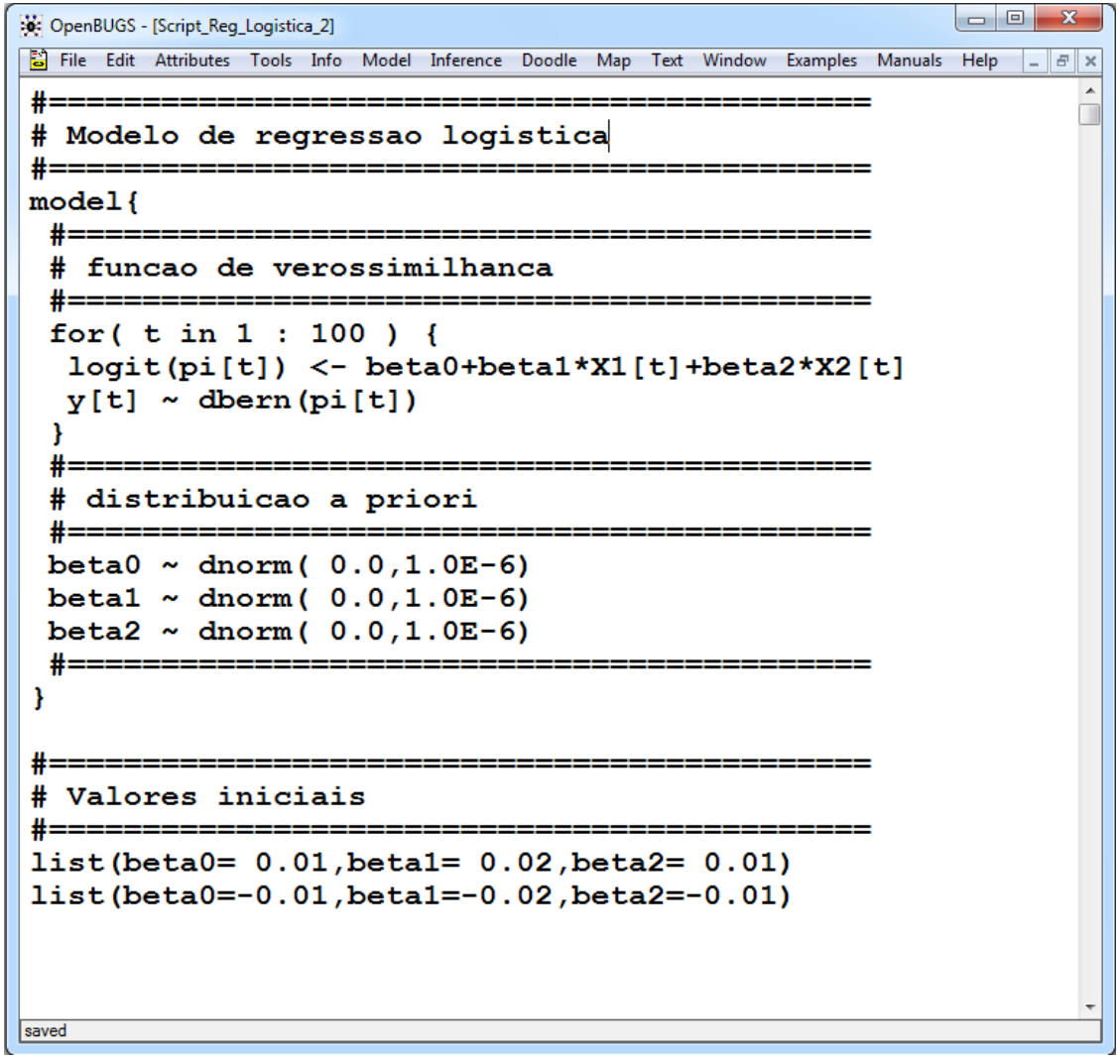
em que $f(\boldsymbol{\beta})$ é a distribuição a priori.

Dados artificiais gerados do seguinte modelo:

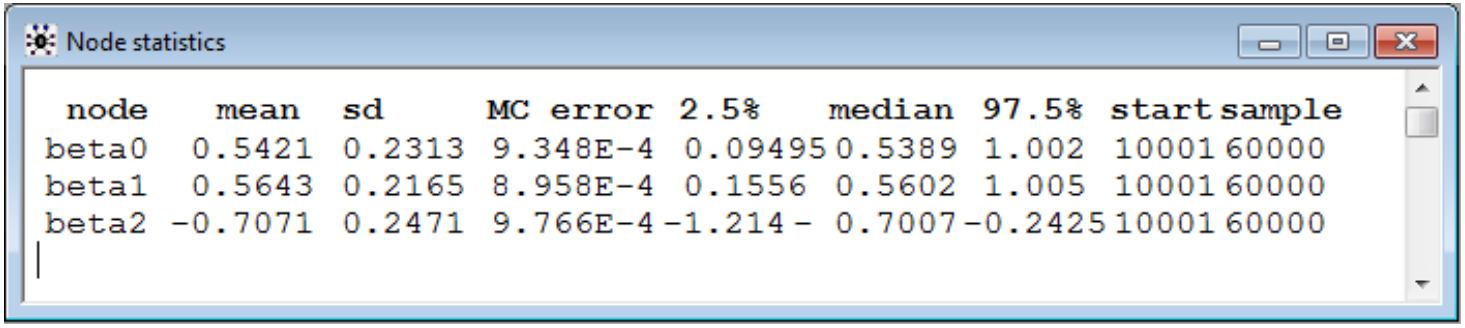
$$\begin{aligned} f(y_t = \omega|\boldsymbol{\beta}) &= \pi_t^{\omega}(1-\pi_t)^{1-\omega}, \quad \omega = 0, 1 \\ \text{logit}(\pi_t) &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t}, \quad t = 1, 2, \ldots, n, \end{aligned}$$

com $\beta_0 = 0,3$, $\beta_1 = 0,5$, $\beta_2 = -0,7$ e $n = 100$.

O modelo em código BUGS



Resultados



Parâmetro	Média	D.P.	2,5%	Mediana	97,5%
β_0	0,5421	0,2313	0,095	0,539	1,002
β_1	0,5643	0,2165	0,156	0,560	1,005
β_2	-0,7071	0,2471	-1,214	-0,701	-0,243

Ver arquivos regressao_logistica.odc e dados_regressao_logistica.txt

Exemplo

Considere o seguinte modelo de volatilidade estocástica:

$$\begin{aligned} y_t &= \exp\{h_t/2\}\varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0,1) \\ h_t &= \alpha + \phi(h_{t-1} - \alpha) + \omega_t, & \omega_t &\sim \mathcal{N}(0,\sigma^2), \end{aligned}$$

para $t = 1, 2, \ldots, T$. Temos que

- y_t é o retorno composto, $y_t = \log(P_t) - \log(P_{t-1})$;
- h_t é a log-volatilidade;
- α é o intercepto (constante);
- ϕ é o parâmetro de persistência; e
- σ^2 é a variância das log-volatilidades.

A distribuição a posteriori é dada por

$$\begin{aligned} f(\alpha, \phi, \sigma^2, \mathbf{h_{1:T}}|\mathbf{y}) \quad &\propto \quad \left[\prod_{t=1}^{T} f(y_t|h_t)\right] \left[\prod_{t=2}^{T} f(h_t|h_{t-1}, \alpha, \phi, \sigma^2)\right] \\ &\times \quad f(h_1|\alpha, \phi, \sigma^2) f(\alpha, \phi, \sigma^2), \end{aligned}$$

em que $f(\alpha, \phi, \sigma^2)$ é a distribuição a priori.

Em geral, assumimos que $f(\alpha, \phi, \sigma^2) = f(\alpha) f(\phi) f(\sigma^2)$.

Temos a restrição $|\phi| < 1$. Vamos impor que $0 \leqslant \phi < 1$.

A distribuição a priori é dada por:

$$\begin{aligned} h_1 &\sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{1-\phi^2}\right) \\ \alpha &\sim \mathcal{N}(0; 0,0001) \\ \phi &\sim \text{Beta}(20; 1, 5) \\ \sigma^2 &\sim \mathcal{IG}(2,5; 0,025). \end{aligned}$$

No algoritmo de Gibbs, temos que

$$\begin{aligned} (\alpha|\phi, \sigma^2, \mathbf{h_{1:T}}, \mathbf{y}) &\sim \mathcal{N}(a_\alpha, b_\alpha^2); \\ (\phi|\alpha, \sigma^2, \mathbf{h_{1:T}}, \mathbf{y}) &\sim \text{Desconhecida, Metropolis-Hastings;} \\ (\sigma^2|\alpha, \phi, , \mathbf{h_{1:T}}, \mathbf{y}) &\sim \mathcal{IG}(c_\sigma, d_\sigma); \quad \text{e} \\ (\mathbf{h_{1:T}}|\alpha, \phi, \sigma^2, \mathbf{y}) &\sim \text{Desconhecida, Metropolis-Hastings.} \end{aligned}$$

Comentários:

- Gerar uma amostra da condicional de $\mathbf{h_{1:T}}$ pode ser feito usando o OpenBUGS.
- Você não precisa programar nada, porém a convergência pode ser lenta devido a amostragem de cada um dos h_t ser feita separadamente.
- Para evitar autocorrelações altas dos sucessivos estados podemos gerar uma amostra da condicional de $\mathbf{h_{1:T}}$ usando um algoritmo em blocos.
- Isto melhora a convergência do algoritmo, porém você tem que escrever o código em uma linguagem de programação.

Retornos do Reino Unido

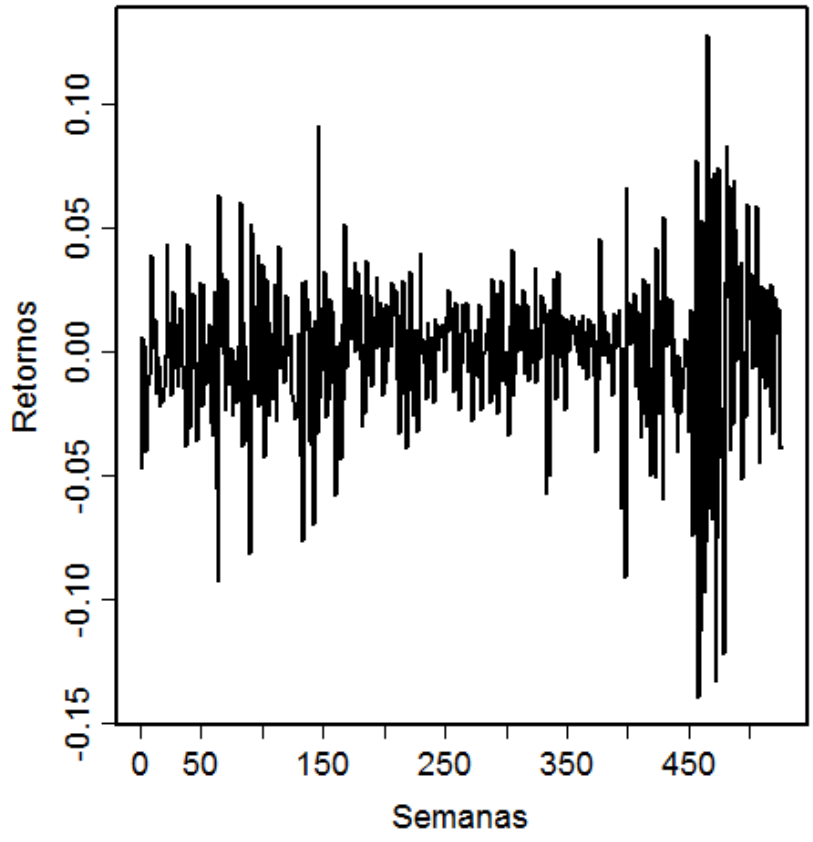


Figura 15: Retornos semanais do Reino Unido de 6/jan/2000 a 28/jan/2010 (T=526).

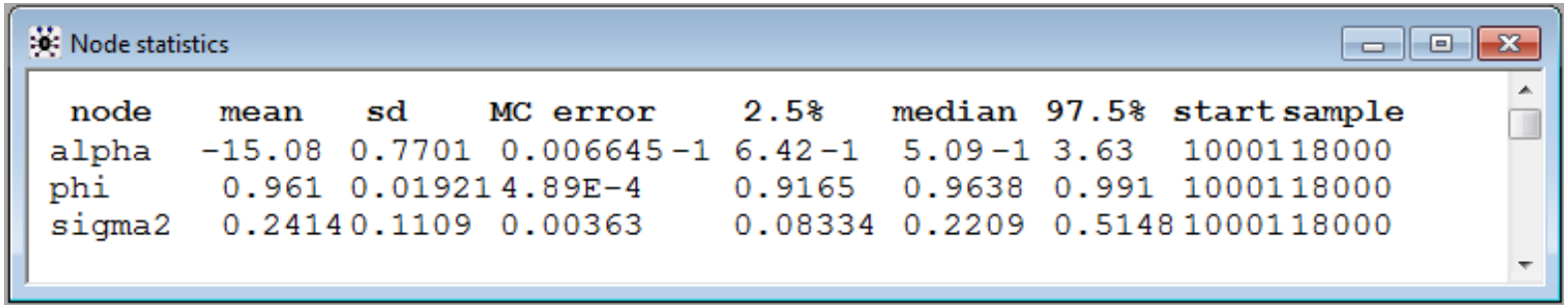
O modelo em código BUGS

volatilidade_estocastica

```
#============================================
# Modelo de volatilidade estocastica
#============================================
model{
 for(t in 1:526){
  eta[t] <- exp(-h[t])
  y[t]   ~ dnorm(0.0,eta[t])
 }
 for(t in 2:526){
  mu[t] <- alpha + phi*(h[t-1]-alpha)
  h[t]  ~  dnorm(mu[t],tau)
 }
 h[1]    ~  dnorm(mu[1],tau0)
 alpha   ~  dnorm(0.0,0.0001)
 phi     ~  dbeta(2.5,1.5)
 tau     ~  dgamma(2.5,0.025)
 sigma2 <- 1/tau
 mu[1]  <- alpha
 tau0   <- tau*(1-pow(phi,2))
}

#============================================
# Valores iniciais
#============================================
list(alpha=0.0,phi=0.95,tau=1.0)
list(alpha=-0.1,phi=0.99,tau=0.5)
```

Resultados



Parâmetro	Média	D.P.	2,5%	Mediana	97,5%
α	-15,08	0,7701	-16,42	-15.09	-13,63
ϕ	0,961	0,0192	0,917	0,963	0,991
σ^2	0,241	0,1109	0,083	0,221	0,515

Ver arquivos volatilidade_estocastica.odc e dados_UK.txt

Programas alternativos

Podemos utilizar o pacote **R2OpenBUGS** do programa **R**.

Ver arquivo exemplo_60.r

Uma alternativa é o **JAGS**, disponível no sítio mcmc-jags.sourceforge.io.

Podemos utilizar o pacote **rjags** do programa **R**.

Ver arquivo exemplo_61.r

Outra alternativa é o **Stan**, disponível no sítio mc-stan.org.

Podemos utilizar o pacote **rstan** do programa **R**.

Escolha de modelos

Suponha que existam pelo menos dois modelos paramétricos para um mesmo conjunto de dados $(y_1, \ldots, y_n)$.

Por exemplo,

- $(y_i|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$, independentes, $i = 1, 2, \ldots, n$; ou
- $(y_i|\mu, \sigma^2, \nu) \sim t_\nu(\mu, \sigma^2)$, independentes, $i = 1, 2, \ldots, n$; ou
- $(y_i|\mu, \tau) \sim \mathcal{LP}(\mu, \tau)^2$, independentes, $i = 1, 2, \ldots, n$.

Nestes casos, as inferências estariam completas quando obtivermos cada uma das distribuições a posteriori dos parâmetros desconhecidos.

Contudo, qual seria o modelo mais adequado para os dados?

[2] Laplace

Introdução

Considere um modelo para um conjunto de dados $\boldsymbol{y}$. Podemos calcular a função (de densidade) de probabilidade **preditiva** dada por

$$f(\boldsymbol{y}|\mathcal{M}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}) f(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}.$$

Se $\boldsymbol{\theta}$ tiver componentes discretas, então a integral se transforma em somatórios adequados para estas variáveis.

Note que $f(\boldsymbol{y}|\mathcal{M})$ é a constante de normalização da densidade a posteriori de $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M})$.

A função (de densidade) de probabilidade preditiva pode ser vista como a **"função de verossimilhança"** do modelo $\mathcal{M}$.

Portanto, pode ser utilizada para fazer inferência sobre a escolha de um modelo.

Inferência para o modelo

Considere o problema geral de escolher um modelo dentre J possíveis dado um conjunto de dados observados $\boldsymbol{y}$.

Atribua probabilidades a priori para cada modelo possível:

$$f(\mathcal{M}_1), \ldots, f(\mathcal{M}_J),$$

tal que $\sum_{j=1}^{J} f(\mathcal{M}_j) = 1$.

Uma abordagem direta é estimar a probabilidade a posteriori dos modelo. Nesse caso,

$$f(\mathcal{M}_j|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\mathcal{M}_j)f(\mathcal{M}_j)}{\sum_{i=1}^{J} f(\boldsymbol{y}|\mathcal{M}_i)f(\mathcal{M}_i)},$$

em que $f(\boldsymbol{y}|\mathcal{M}_j)$ é a função de verossimilhança marginal para os dados sob o modelo $\mathcal{M}_j$.

Cálculo da preditiva

Para o modelo $\mathcal{M}_j$, nosso principal interesse é calcular

$$f(\boldsymbol{y}|\mathcal{M}_j) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}_j) f(\boldsymbol{\theta}|\mathcal{M}_j) d\boldsymbol{\theta}.$$

Em muitos casos, a expressão da preditiva não pode ser obtida analiticamente devido a complexidade da integral acima.

Nestes casos, usamos métodos aproximados para obter a preditiva do modelo $\mathcal{M}_j$.

Consideraremos o método de **Monte Carlo** para aproximar a preditiva de interesse.

Em alguns casos, utilizaremos amostras da posteriori de $\boldsymbol{\theta}$ obtidas usando Monte Carlo via cadeias de Markov para obter aproximações da preditiva de $\boldsymbol{y}$.

Aproximação usando a distribuição a priori

Considere um modelo $\mathcal{M}$. Temos que a preditiva é dada por

$$f(\boldsymbol{y}|\mathcal{M}) = \int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}) f(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}.$$

Note que podemos escrever

$$f(\boldsymbol{y}|\mathcal{M}) = \mathrm{E}\left(f(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})\right),$$

sob a distribuição a priori de $\boldsymbol{\theta}$.

Aproximação baseada na distribuição a priori:

$$\widehat{f}_1(\boldsymbol{y}|\mathcal{M}) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)}, \mathcal{M}),$$

em que $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ são gerados da distribuição a priori $f(\boldsymbol{\theta}|\mathcal{M})$.

Exemplo

Considere $y_1, \ldots, y_n$ com distribuição exponencial com parâmetro θ,

$$f(y_i|\theta) = \theta \exp\{-\theta y_i\}, \quad \text{para} \quad y_i > 0.$$

Para a distribuição a priori $\theta \sim \mathcal{G}(a, b)$, temos que a preditiva é dada por

$$f(\boldsymbol{y}) = \frac{\Gamma(n+a)}{\Gamma(a)} \frac{b^a}{(b+n\overline{y})^{n+a}}.$$

Podemos usar a aproximação baseada em gerações da distribuição a priori:

- Gerar $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ da $\mathcal{G}(a, b)$.
- Calcular $L(\theta^{(k)}|\boldsymbol{z}) = (\theta^{(k)})^n \exp\{-\theta^{(k)} n\overline{y}\}$, para $k = 1, \ldots, N$.
- Calcular $\widehat{f}_1(\boldsymbol{y}) = \dfrac{1}{N} \displaystyle\sum_{k=1}^{N} L(\theta^{(k)}|\boldsymbol{y})$.

Ver arquivo exemplo_62.r

Estimador média harmônica

$$
\begin{aligned}
1 &= \int_{\Theta} \frac{f(\boldsymbol{y}|\mathcal{M})}{f(\boldsymbol{y}|\boldsymbol{\theta},\mathcal{M})} \times \frac{f(\boldsymbol{y}|\boldsymbol{\theta},\mathcal{M})}{f(\boldsymbol{y}|\mathcal{M})} \times f(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \\
&= f(\boldsymbol{y}|\mathcal{M}) \int_{\Theta} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta},\mathcal{M})} f(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M}) d\boldsymbol{\theta} \quad \Rightarrow \\
f(\boldsymbol{y}|\mathcal{M}) &= \left[\int_{\Theta} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta},\mathcal{M})} f(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{M}) d\boldsymbol{\theta} \right]^{-1}
\end{aligned}
$$

O **estimador de média harmônica** é dado por

$$
\widehat{f}_2(\boldsymbol{y}|\mathcal{M}) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)},\mathcal{M})} \right]^{-1},
$$

em que $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ são gerados da posteriori de $\boldsymbol{\theta}$.

Note que um valor muito pequeno da função de verossimilhança tem grande efeito sobre o estimador, tornando-o muito instável.

Ver arquivo exemplo_63.r

Amostragem por importância

Estimador de Monte Carlo para I:

$$\widehat{I}_1 = \frac{1}{N}\sum_{i=1}^{N} h(x_i)\frac{f(x_i)}{g(x_i)},$$

Considere uma alternativa que garante variância finita do estimador $\widehat{I}$:

$$\widehat{I}_2 = \frac{\sum_{i=1}^{N} h(x_i)\frac{f(x_i)}{g(x_i)}}{\sum_{i=1}^{N}\frac{f(x_i)}{g(x_i)}}.$$

Note que, neste caso, estamos substituindo N por $\sum_{i=1}^{N} f(x_i)/g(x_i)$, que é a soma dos pesos.

Como $\sum_{i=1}^{N} f(x_i)/g(x_i) \to 1$ quando $N \to \infty$ o estimador $\widehat{I}_2 \to I$ pela Lei Forte dos Grandes Números.

Outro estimador

Os problemas de $\widehat{f}_1$ e $\widehat{f}_2$ são "opostos", de forma que uma solução é considerar uma mistura das duas propostas.

Omitindo a dependência sobre o modelo $\mathcal{M}$, temos que

$$\widehat{f}_3(\boldsymbol{y}) = \frac{\sum_{i=1}^{N} f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})\omega(\boldsymbol{\theta}^{(i)})}{\sum_{i=1}^{N} \omega(\boldsymbol{\theta}^{(i)})} = \frac{\sum_{i=1}^{N} f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)}) \left[\delta f(\boldsymbol{y}) + (1-\delta) f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})\right]^{-1}}{\sum_{i=1}^{N} \left[\delta f(\boldsymbol{y}) + (1-\delta) f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})\right]^{-1}},$$

em que

$$\omega(\boldsymbol{\theta}^{(i)}) = \frac{f(\boldsymbol{\theta}^{(i)})}{\delta f(\boldsymbol{\theta}^{(i)}) + (1-\delta) f(\boldsymbol{\theta}^{(i)}|\boldsymbol{y})},$$

δ é o peso da mistura e deve ser pequeno, e $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ são gerados da mistura $\delta f(\boldsymbol{\theta}) + (1-\delta) f(\boldsymbol{\theta}^{(i)}|\boldsymbol{y})$.

Note que o estimador depende de $f(\boldsymbol{y})$ que é desconhecido. Solução?

Algoritmo iterativo

1. Inicializar $\widehat{f}_3^{(0)}$.
2. Gerar δN valores da distribuição a priori e gerar $(1-\delta)N$ valores da distribuição a posteriori.
3. Para j de 1 até J faça

$$\widehat{f}_3^{(j)}(\boldsymbol{y}) = \frac{\sum_{i=1}^{N} f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)}) \left[\delta \widehat{f}_3^{(j-1)}(\boldsymbol{y}) + (1-\delta) f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})\right]^{-1}}{\sum_{i=1}^{N} \left[\delta \widehat{f}_3^{(j-1)}(\boldsymbol{z}) + (1-\delta) f(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})\right]^{-1}}.$$

O algoritmo termina para um número máximo de iterações J.

Note que precisamos gerar amostras da priori e da posteriori.

Ver arquivo exemplo_64.r

Amostragem por ponte *(bridge sampling)*

Considere uma função $\alpha(\boldsymbol{\theta})$, chamada "ponte". Daí, temos a relação:

$$\int \alpha(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \frac{f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})} d\boldsymbol{\theta} = \int \alpha(\boldsymbol{\theta}) g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

Então,

$$f(\boldsymbol{y}) = \frac{\int \alpha(\boldsymbol{\theta}) f(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \alpha(\boldsymbol{\theta}) g(\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}}.$$

Considere $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N_1)})$ uma amostra da posteriori e $(\widetilde{\boldsymbol{\theta}}^{(1)}, \ldots, \widetilde{\boldsymbol{\theta}}^{(N_2)})$ uma amostra de $g(\boldsymbol{\theta})$.

O estimador por "ponte" (*bridge sampling*) é dado por

$$\widehat{f}_4(\boldsymbol{y}) = \frac{\frac{1}{N_2} \sum_{j=1}^{N_2} \alpha(\widetilde{\boldsymbol{\theta}}^{(j)}) f(\widetilde{\boldsymbol{\theta}}^{(j)}) f(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}^{(j)})}{\frac{1}{N_1} \sum_{j=1}^{N_1} \alpha(\boldsymbol{\theta}^{(j)}) g(\boldsymbol{\theta}^{(j)})}.$$

Estimador por ponte ótimo

O estimador que minimiza o erro quadrático médio (EQM) tem

$$\alpha(\boldsymbol{\theta}) = \frac{N_2}{N_1 f(\boldsymbol{\theta}|\boldsymbol{y}) + N_2 g(\boldsymbol{\theta})}.$$

Defina $s_1 = N_1/(N_1 + N_2)$ e $s_2 = N_2/(N_1 + N_2)$. Defina

$$\omega_j = \frac{f(\boldsymbol{\theta}^{(j)}) f(\boldsymbol{y}|\boldsymbol{\theta}^{(j)})}{g(\boldsymbol{\theta}^{(j)})} \quad \text{e} \quad \widetilde{\omega}_j = \frac{f(\widetilde{\boldsymbol{\theta}}^{(j)}) f(\boldsymbol{y}|\widetilde{\boldsymbol{\theta}}^{(j)})}{g(\widetilde{\boldsymbol{\theta}}^{(j)})}.$$

O estimador com erro quadrático médio mínimo é dado por

$$\widehat{f}_4(\boldsymbol{y}) = \frac{\dfrac{1}{N_2}\displaystyle\sum_{j=1}^{N_2} \widetilde{\omega}_j \left[s_1 \widetilde{\omega}_j + s_2 \widehat{f}(\boldsymbol{y})\right]^{-1}}{\dfrac{1}{N_1}\displaystyle\sum_{j=1}^{N_1} \left[s_1 \omega_j + s_2 \widehat{f}(\boldsymbol{y})\right]^{-1}},$$

e estima $f(\boldsymbol{y})$.

Ver arquivo exemplo_65.r

Estimador gama deslocado

Outra proposta é conhecida como “estimador gama deslocada”.

Nessa proposta, as saídas de Monte Carlo (via cadeias de Markov) são utilizadas para calcular a sequência de valores da log-verossimilhança $\{\ell_k : \ k = 1, \ldots, n\}$ e a distribuição a posteriori das log-verossimilhanças é dada por

$$\ell_{max} - \ell_k \sim \mathcal{G}(\alpha, \lambda),$$

em que ℓ_{max} é o máximo da log-verossimilhança, $\alpha = d/2$ com d sendo o número de parâmetros do modelo e $\lambda < 1$. Na prática, λ é próximo de 1.

Combinando a identidade da média harmônica

$$\frac{1}{f(\boldsymbol{y})} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(\frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta})}\right),$$

com a distribuição gama para $\ell_{max} - \ell_k$, temos

$$\log f(\boldsymbol{y}) = \ell_{max} + \alpha \log(1 - \lambda).$$

Em geral, ℓ_{max} não é conhecido, então

$$\widehat{\ell}_{max} = \max\{\bar{\ell} + s^2_\ell, \ell_k\},$$

é utilizado, em que $\bar{\ell} + s^2_\ell$ é o estimador de momentos de ℓ_{max}, e $\bar{\ell}$ e s^2_ℓ são a média e variância amostrais de ℓ_k, respectivamente.

Assim, temos que

$$\widehat{f}_5(\boldsymbol{y}) = \exp\left\{\widehat{\ell}_{max} + \alpha \log(1 - \lambda)\right\}.$$

Ver arquivo exemplo_66.r

Fator de Bayes

O problema de escolher modelos também pode ser visto como um problema de testar hipóteses.

Por exemplo, podemos usar o Fator de Bayes para medir a probabilidade a posteriori relativa dos modelos de interesse.

Se consideramos dois modelos $\mathcal{M}_1$ e $\mathcal{M}_0$, então o Fator de Bayes é definido por

$$B_{10} = \frac{f(\boldsymbol{y}|\mathcal{M}_1)}{f(\boldsymbol{y}|\mathcal{M}_0)}.$$

O Fator de Bayes resume a evidência fornecida pelos dados em favor de um modelo contra o outro.

É usual considerar 2 vezes o log do Fator de Bayes, pois nesse caso temos a mesma escala da estatística do teste da razão de funções de verossimilhanças.

Fator de Bayes na prática

Suponha que

$$\begin{cases} H_0: & \text{o modelo escolhido deve ser o } \mathcal{M}_0; \\ H_1: & \text{o modelo escolhido deve ser o } \mathcal{M}_1. \end{cases}$$

Abaixo, segue um guia para interpretação do Fator de Bayes B_{10} que é a evidência em favor do modelo $\mathcal{M}_1$ contra o modelo $\mathcal{M}_0$.

$2\log(B_{10})$	B_{10}	Evidência contra H_0
0 a 2	1 a 3	Não merece ser muito comentada
2 a 6	3 a 20	Positiva
6 a 10	20 a 150	Forte
>10	>150	Muito forte

Exemplo

Suponha que queremos comparar 2 modelos:

$$\begin{aligned} \mathcal{M}_0 : &\qquad y \sim \mathcal{LN}(\mu, \sigma^2) \\ \mathcal{M}_1 : &\qquad y \sim \mathcal{E}(\theta). \end{aligned}$$

Podemos calcular a preditiva $f(\boldsymbol{y})$ para cada modelo e obter o fator de Bayes.

Ver arquivo exemplo_67.r

Modelos com dimensão variável

Definição

Um modelo com dimensão variável é um modelo em que uma das quantidades desconhecidas é o número de quantidades desconhecidas.

Em outras palavras, a dimensão do espaço de parâmetros não é fixa.

Isso está relacionado com o problema de seleção de modelos.

Exemplo

Considere K covariáveis que possivelmente estão relacionadas com a variável resposta y, em um modelo de regressão,

$$y_i = \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i.$$

Suponha que K seja grande.

Neste caso, temos 2^K modelos possíveis.

Como selecionar modelos nesse contexto?

Seleção ou estimação?

Considere o problema geral de escolher entre K modelos para um conjunto de dados observados $\boldsymbol{y}$,

$$\mathcal{M}_k = \{f(\boldsymbol{y}|\boldsymbol{\theta}_k) : \ \boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k\}.$$

Note que o espaço paramétrico em um problema com K modelos é

$$\boldsymbol{\Theta} = \bigcup_k \Big[\{k\} \times \boldsymbol{\Theta}_k\Big].$$

Estimação nesse contexto não é um problema trivial.

Por exemplo, ao considerarmos esse espaço paramétrico há uma tendência a sobrestimar o número de parâmetros levando a um sobreajuste dos dados.

Por outro lado, se consideramos esse um problema puramente de decisão, escolheremos um modelo mais provável, que pode levar a grandes erros.

Seleção bayesiana de modelos

Atribua probabilidades a priori para cada modelo possível:

$$f(\mathcal{M}_1), \ldots, f(\mathcal{M}_K),$$

tal que $\sum_{k=1}^{K} f(\mathcal{M}_k) = 1$.

E também uma distribuição a priori para os parâmetros de cada modelo:

$$f(\boldsymbol{\theta}_k|\mathcal{M}_k), \quad k = 1, \ldots, K.$$

Abordagem direta: estimar a probabilidade a posteriori de cada modelo.

Nesse caso,

$$f(\mathcal{M}_k|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\mathcal{M}_k)f(\mathcal{M}_k)}{\sum_{j=1}^{K} f(\boldsymbol{y}|\mathcal{M}_j)f(\mathcal{M}_j)},$$

em que $f(\boldsymbol{y}|\mathcal{M}_k)$ é a preditiva do modelo $\mathcal{M}_k$.

Reescrevendo, temos que

$$f(\mathcal{M}_k|\boldsymbol{y}) = \frac{f(\mathcal{M}_k)\int_{\Theta_k} f(\boldsymbol{y}|\boldsymbol{\theta}_k,\mathcal{M}_k)f(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k}{\sum_{j=1}^{K} f(\mathcal{M}_j)\int_{\Theta_j} f(\boldsymbol{y}|\boldsymbol{\theta}_j,\mathcal{M}_j)f(\boldsymbol{\theta}_j|\mathcal{M}_j)d\boldsymbol{\theta}_j},$$

a probabilidade a posteriori do modelo $\mathcal{M}_k$.

Decisão: escolher o modelo com a maior probabilidade $f(\mathcal{M}_k|\boldsymbol{y})$ ou uma combinação de modelos.

Seja z a observação a ser predita,

$$f(z|\mathcal{M}_k,\boldsymbol{y}) = \int_{\Theta_k} f(z|\boldsymbol{\theta}_k,\boldsymbol{y},\mathcal{M}_k)f(\boldsymbol{\theta}_k|\boldsymbol{y},\mathcal{M}_k)d\boldsymbol{\theta}_k,$$

e, portanto,

$$f(z|\boldsymbol{y}) = \sum_{k=1}^{K} f(z|\mathcal{M}_k,\boldsymbol{y})f(\mathcal{M}_k|\boldsymbol{y}).$$

Medidas para seleção de modelos

Vimos que o fator de Bayes pode ser usado para comparar vários modelos,

$$B_{10} = \frac{f(\boldsymbol{y}|\mathcal{M}_1)}{f(\boldsymbol{y}|\mathcal{M}_0)}.$$

Reescrevendo, temos que a chance a posteriori é dada por

$$\frac{f(\mathcal{M}_1|\boldsymbol{y})}{f(\mathcal{M}_0|\boldsymbol{y})} = \frac{f(\mathcal{M}_1)f(\boldsymbol{y}|\mathcal{M}_1)}{f(\mathcal{M}_0)f(\boldsymbol{y}|\mathcal{M}_0)} = \frac{f(\mathcal{M}_1)}{f(\mathcal{M}_0)} \times B_{10}.$$

Critério preditivo a posteriori (EPD)

Sob perda quadrática, devemos escolher o modelo que minimiza

$$\mathrm{E}([\boldsymbol{z}-\boldsymbol{y}]^\top[\boldsymbol{z}-\boldsymbol{y}]|\boldsymbol{y},\mathcal{M}_j)=\sum_{i=1}^{n}\mathrm{Var}(z_i|\boldsymbol{y},\mathcal{M}_j)+\sum_{i=1}^{n}[y_i-\mathrm{E}(z_i|\boldsymbol{y},\mathcal{M}_j)]^2\,,$$

em que $z_1,\ldots,z_n$ são previsões (ou réplicas) de $y_1,\ldots,y_n$.

Definimos

$$\begin{aligned} EPD &= P+G \quad \text{(penalidade + ajuste)} \quad \text{com} \\ P &= \sum_{i=1}^{n}\mathrm{Var}(z_i|\boldsymbol{y},\mathcal{M}_j) \quad \text{e} \quad G = \sum_{i=1}^{n}[y_i-\mathrm{E}(z_i|\boldsymbol{y},\mathcal{M}_j)]^2. \end{aligned}$$

Sejam $(z_i^{(1)},z_i^{(2)},\ldots,z_i^{(M)})$ réplicas (de Monte Carlo) de y_i. Então, podemos aproximar

$$\mathrm{E}(z_i|\boldsymbol{y},\mathcal{M}_j)\simeq\overline{z}_i=\frac{1}{M}\sum_{k=1}^{M}z_i^{(k)};\;\;\mathrm{Var}(z_i|\boldsymbol{y},\mathcal{M}_j)\simeq\frac{1}{M-1}\sum_{k=1}^{M}(z_i^{(k)}-\overline{z}_i)^2.$$

Ver arquivo exemplo_68.r

Critério de informação do desvio (DIC, do inglês)

Sejam $\boldsymbol{\theta}^\star = \mathrm{E}(\boldsymbol{\theta}|\boldsymbol{y})$ e $D(\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta})$. Então,

$$DIC = \overline{D} + p_d \quad \text{(qualidade de ajuste + penalização)},$$

em que

$$\overline{D} = \mathrm{E}(D(\boldsymbol{\theta})|\boldsymbol{y}) \quad \text{e} \quad p_d = \overline{D} - D(\boldsymbol{\theta}^\star).$$

O menor DIC indica o melhor ajuste do modelo.

O DIC é computacionalmente atrativo pois pode ser facilmente calculado das saídas de um Monte Carlo via cadeias de Markov.

Sejam $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ valores gerados utilizando Monte Carlo via cadeias de Markov. Então,

$$\overline{D} = \mathrm{E}(D(\boldsymbol{\theta})|\boldsymbol{y}) \simeq \frac{1}{M}\sum_{k=1}^{M} D(\boldsymbol{\theta}^{(k)})$$

e também

$$D(\boldsymbol{\theta}^\star) \simeq D(\overline{\boldsymbol{\theta}}), \quad \text{com} \quad \overline{\boldsymbol{\theta}} = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{\theta}^{(k)}.$$

Ver arquivos regressao.odc e dados_regressao.txt

Monte Carlo via cadeias de Markov com saltos reversíveis

Considere o problema de fazer inferência para dois modelos e seus parâmetros simultaneamente.

Considere que $\mathcal{M}_1$ tem um parâmetro θ e $\mathcal{M}_2$ tem dois parâmetros $\{\theta_1, \theta_2\}$.

O algoritmo de Monte Carlo via cadeias de Markov com saltos reversíveis faz inferência para θ, θ_1 e θ_2, e para os modelos $\mathcal{M}_1$ e $\mathcal{M}_2$.

Em um único algoritmo iremos mover a cadeia do modelo $\mathcal{M}_1$ para o $\mathcal{M}_2$ e do modelo $\mathcal{M}_2$ para o $\mathcal{M}_1$.

O movimento de $\mathcal{M}_2$ para $\mathcal{M}_1$ é determinístico. Por exemplo, $\theta = \dfrac{\theta_1 + \theta_2}{2}$.

E o movimento de $\mathcal{M}_1$ para $\mathcal{M}_2$?

Uma maneira interessante de simular ambos parâmetros e modelos é utilizando Monte Carlo via cadeias de Markov com saltos reversíveis.

O algoritmo é construído de forma a manter a condição de "balanceada detalhada".

Por exemplo, considere $\dim(\boldsymbol{\Theta}_1) > \dim(\boldsymbol{\Theta}_2)$.

Então, o movimento de $\boldsymbol{\Theta}_1$ para $\boldsymbol{\Theta}_2$ pode ser representado por uma transformação determinística de $\boldsymbol{\theta}_1$,

$$\boldsymbol{\theta}_2 = T(\boldsymbol{\theta}_1).$$

Impõe-se uma condição de igualdade de dimensões em que o movimento oposto de $\boldsymbol{\Theta}_2$ para $\boldsymbol{\Theta}_1$ é concentrado em $\{\boldsymbol{\theta}_1 : \ \boldsymbol{\theta}_2 = T(\boldsymbol{\theta}_1)\}$.

No caso geral, se $\boldsymbol{\theta}_1$ é completado por simulação $\boldsymbol{u}_1 \sim g_1(\boldsymbol{u}_1)$ em $(\boldsymbol{\theta}_1, \boldsymbol{u}_1)$ e $\boldsymbol{\theta}_2$ por $\boldsymbol{u}_2 \sim g_2(\boldsymbol{u}_2)$ em $(\boldsymbol{\theta}_2, \boldsymbol{u}_2)$ tal que o mapeamento entre $(\boldsymbol{\theta}_1, \boldsymbol{u}_1)$ e $(\boldsymbol{\theta}_2, \boldsymbol{u}_2)$ é uma bijeção,

$$(\boldsymbol{\theta}_2, \boldsymbol{u}_2) = T(\boldsymbol{\theta}_1, \boldsymbol{u}_1).$$

O algoritmo se basea na seguinte identidade:

$$f(\boldsymbol{\theta}_k, k|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_k, k) f_k(\boldsymbol{\theta}_k|k) f(k)}{f(\boldsymbol{y})}.$$

A probabilidade de aceitar um movimento de $\mathcal{M}_1$ para $\mathcal{M}_2$ é dada por

$$\min\left\{1, \frac{f(\boldsymbol{\theta}_2, 2|\boldsymbol{y})\pi_{21}\phi_{21}(\boldsymbol{u}_2)}{f(\boldsymbol{\theta}_1, 1|\boldsymbol{y})\pi_{12}\phi_{12}(\boldsymbol{u}_1)}\left|\frac{\partial T_{12}(\boldsymbol{\theta}_1, \boldsymbol{u}_1)}{\partial(\boldsymbol{\theta}_1, \boldsymbol{u}_1)}\right|\right\},$$

em que

- $f(\boldsymbol{\theta}_k, k|\boldsymbol{y})$ é a distribuição a posteriori de $\boldsymbol{\theta}_k$ sob o modelo $\mathcal{M}_k$;
- π_{ij} é a probabilidade de mover do modelo $\mathcal{M}_i$ para o modelo $\mathcal{M}_j$; e
- $\left|\frac{\partial T_{12}(\boldsymbol{\theta}_1, \boldsymbol{u}_1)}{\partial(\boldsymbol{\theta}_1, \boldsymbol{u}_1)}\right|$ é o Jacobiano da transformação.

Algoritmo

Na iteração ℓ, temos $(\boldsymbol{\theta}_k^{(\ell)}, k)$

1. Selecione o modelo $\mathcal{M}_j$ com probabilidade π_{kj};
2. Gere $\boldsymbol{u}_{kj} \sim \phi_{kj}(\boldsymbol{u})$;
3. Defina $(\boldsymbol{\theta}_j^\star, \boldsymbol{v}_{kj}) = T_{kj}(\boldsymbol{\theta}_k^{(\ell)}, \boldsymbol{u}_{kj})$;
4. Tome $\boldsymbol{\theta}_j^{(\ell+1)} = \boldsymbol{\theta}_j^\star$ com probabilidade

$$\min\left\{1, \frac{f(\boldsymbol{\theta}_j^\star, j|\boldsymbol{y})\pi_{jk}\phi_{jk}(\boldsymbol{v}_{jk})}{f(\boldsymbol{\theta}_k^{(\ell)}, k|\boldsymbol{y})\pi_{kj}\phi_{kj}(\boldsymbol{u}_{kj})}\left|\frac{\partial T_{kj}(\boldsymbol{\theta}_k^{(\ell)}, \boldsymbol{u}_{kj})}{\partial(\boldsymbol{\theta}_k^{(\ell)}, \boldsymbol{u}_{kj})}\right|\right\},$$

e tome $\boldsymbol{\theta}_j^{(\ell+1)} = \boldsymbol{\theta}_k^{(\ell)}$ caso contrário.

Escolha da proposta

A escolha das densidades propostas ϕ_{ij} deve ser feita com cautela.

Em especial em altas dimensões.

Amostragem independente: se todos os parâmetros são gerados de sua distribuição proposta, então T_{ij} é identidade e o jacobiano é igual a 1. (Nenhuma transformação é feita nesse caso.)

Se o modelo proposto é igual ao corrente, $\mathcal{M}_k = \mathcal{M}_j$, então o passo de aceitação corresponde a um passo usual de Metropolis.

Proposta independente com momentos baseados na posteriori

Considere que estimativas de média e variância a posteriori estejam disponíveis para os parâmetros de cada modelo.

Nesse caso, uma proposta independente pode ser usada.

Por exemplo,

$$\boldsymbol{\theta}^\star \sim \mathcal{N}(\widehat{\mathrm{E}}(\boldsymbol{\theta}|\boldsymbol{y}), \widehat{\mathrm{Var}}(\boldsymbol{\theta}|\boldsymbol{y})).$$

O movimento de um modelo $\mathcal{M}_k$ para outro $\mathcal{M}_j$ terá jacobiano 1.

A probabilidade de aceitação será dada por

$$\min\left\{1, \frac{\ell_j(\boldsymbol{\theta}_j^\star|\boldsymbol{y}) f_j(\boldsymbol{\theta}_j^\star)\phi_k(\boldsymbol{\theta}_k)}{\ell_k(\boldsymbol{\theta}_k|\boldsymbol{y}) f_k(\boldsymbol{\theta}_k)\phi_j(\boldsymbol{\theta}_j^\star)}\right\}.$$

Exemplo

Falha de ar condicionados de aviões.

Para cada avião, sugere-se usar os modelos abaixo:

$$\begin{aligned} \mathcal{M}_0: \ (y_i|\mu,\sigma^2) &\sim \mathcal{LN}(\mu,\sigma^2), \quad \mu \in \mathbb{R}, \ \sigma^2 > 0 \\ \mathcal{M}_1: \ (y_i|\gamma,\delta) &\sim \text{Weibull}(\gamma,\delta), \ \gamma > 0, \ \delta > 0. \end{aligned}$$

Utilizaremos um Monte Carlo via cadeias de Markov com saltos reversíveis com proposta de transição independente $\boldsymbol{\theta} = (\theta_1, \theta_2)$ tal que $\boldsymbol{\theta} \sim \mathcal{N}_2(\mu_\theta, \boldsymbol{\Sigma}_\theta)$.

Ver arquivo exemplo_69.r

Previsão

Considere $(y_1, y_2, \ldots, y_n)$ uma amostra de uma distribuição com função (de densidade) de probabilidade $f(y|\boldsymbol{\theta})$.

Suponha que estamos interessados em prever um novo valor z.

Uma opção considera a distribuição

$$f(z|\widehat{\boldsymbol{\theta}}, \boldsymbol{y})$$

para algum valor $\widehat{\boldsymbol{\theta}}$ do parâmetro desconhecido.

Note que essa solução ignora a incerteza a respeito de $\boldsymbol{\theta}$.

Solução bayesiana

A solução pela abordagem bayesiana é encontrar a distribuição preditiva

$$\begin{aligned} f(z|\boldsymbol{y}) &= \int_{\Theta} f(z, \boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \\ &= \int_{\Theta} f(z|\boldsymbol{\theta}, \boldsymbol{y}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}. \end{aligned}$$

Essa solução incorpora incerteza a respeito de $\boldsymbol{\theta}$.

Aproximação por Monte Carlo

Note que temos a esperança a posteriori de $g(\boldsymbol{\theta})$ para $g(\boldsymbol{\theta}) = f(z|\boldsymbol{\theta}, \boldsymbol{y})$:

$$f(z|\boldsymbol{y}) = \int_{\Theta} f(z|\boldsymbol{\theta}, \boldsymbol{y}) f(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

Então, o método de Monte Carlo pode ser usado para aproximar a distribuição preditiva de z.

Seja $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})$ uma amostra obtida da distribuição a posteriori de $\boldsymbol{\theta}$.

Uma aproximação para a preditiva é

$$\widehat{f}(z|\boldsymbol{y}) = \frac{1}{M} \sum_{k=1}^{M} f(z|\boldsymbol{\theta}^{(k)}, \boldsymbol{y}).$$

Algoritmo

Obtendo uma amostra de tamanho K da preditiva $f(z|\boldsymbol{y})$:

1. Sortear a componente k do conjunto $\{1, 2, \ldots, M\}$ com probabilidade $1/M$;
2. Gerar z de $f(z|\boldsymbol{y}, \boldsymbol{\theta}^{(k)})$; e
3. Repetir os passos 1 e 2 até obter uma amostra de tamanho K de $f(z|\boldsymbol{y})$.

Dada a amostra de z, poderemos calcular média, variância e percentis amostrais como resumo da distribuição do valor previsto.

Ademais, podemos contruir histogramas como aproximação da função (de densidade) de probabilidade $f(z|\boldsymbol{y})$.

Exemplo

Considere o modelo autoregressivo de ordem 1, AR(1),

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

em que $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, $|\phi| < 1$ e $\sigma^2 > 0$.

Considere a série temporal observada abaixo ($t = 1, 2, \ldots, T$). Como obter a previsão para o tempo $T + 1$?

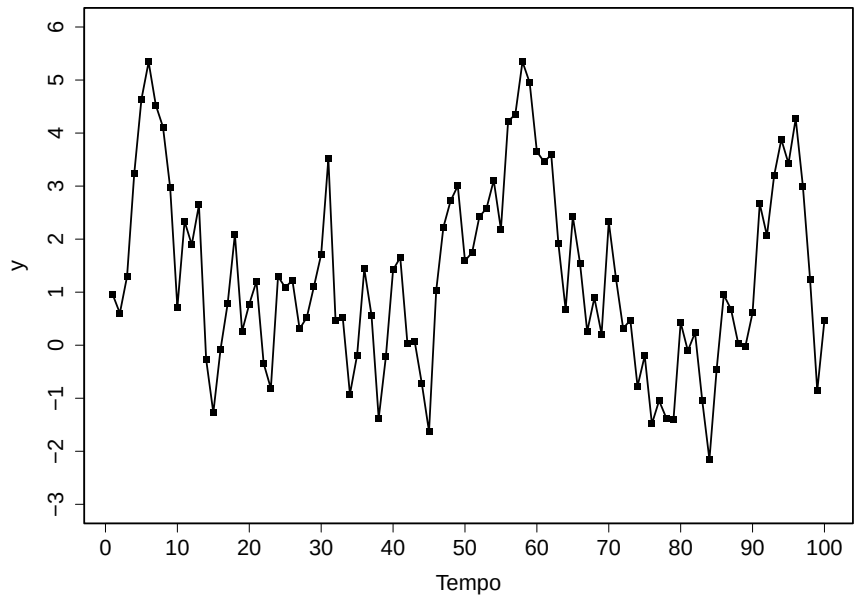


Figura 16: Dados gerados de um processo AR(1).

Função de verossimilhança

A função de verossimilhança (condicional a primeira observação) é dada por

$$L(\phi, \sigma^2) = \prod_{t=2}^{T} f(y_t | y_{t-1}, \phi, \sigma^2).$$

Inferência bayesiana

Considere a distribuição a priori

$$f(\phi, \sigma^2) \propto \sigma^{-2}.$$

As distribuições condicionais completas são dadas por

$$\begin{aligned}
(\sigma^2 | \boldsymbol{y}, \phi) &\sim \mathcal{IG}\left(\frac{T-1}{2}, \frac{\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2}{2}\right) \\
(\phi | \boldsymbol{y}, \sigma^2) &\sim \mathcal{N}\left(\frac{\sum_{t=2}^{T} y_t y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}, \frac{\sigma^2}{\sum_{t=2}^{T} y_{t-1}^2}\right) \boldsymbol{I}(-1 < \phi < 1).
\end{aligned}$$

Suponha que $\phi \in (0, 1)$.

A proposta abaixo considera essa restrição:

$$\begin{aligned} \theta &= \log\left(\frac{\phi}{1-\phi}\right), \quad \theta \in \mathbb{R} \\ \theta^\star &\sim \mathcal{N}(\theta^{(j-1)}, \tau^2), \quad \theta^\star \in \mathbb{R} \\ \phi^\star &= \frac{\exp\{\theta^\star\}}{1+\exp\{\theta^\star\}}, \quad \phi^\star \in (0, 1). \end{aligned}$$

A função de densidade proposta é

$$f(\phi^\star|\phi^{(j-1)}) = f_{\mathcal{N}}(\theta^\star|\theta^{(j-1)}, \tau^2)\phi^\star(1-\phi^\star).$$

Ver arquivo exemplo_70.r

Amostragem composta

Na prática usamos a amostragem composta para gerar de $f(z|\boldsymbol{y})$.

Esse método evita gerar da mistura $\widehat{f}(z|\boldsymbol{y})$.

Algoritmo

Para $i = 1, \ldots, M$,

1. gere $\boldsymbol{\theta}^{(i)}$ de $f(\boldsymbol{\theta}|\boldsymbol{y})$; e
2. gere $z^{(i)}$ de $f(z|\boldsymbol{\theta}^{(i)}, \boldsymbol{y})$.

A amostra $\{z^{(1)}, \ldots, z^{(M)}\}$ é uma amostra de $f(z|\boldsymbol{y})$.

Ver arquivo exemplo_70.r

Dados faltantes

Muitos estudos apresentam dados faltantes.

Seja $\boldsymbol{y}_{falt}$ o dado faltante e $\boldsymbol{y}_{obs}$ o dado observado.

O método bayesiano fornece uma abordagem natural que trata o problema sem usar imputação *ad hoc*.

Suponha que $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{falt})$ tem densidade conjunta $f(\boldsymbol{y}|\boldsymbol{\theta})$.

Sob o ponto de vista bayesiano, temos que

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{y}_{falt}|\boldsymbol{y}_{obs}) &\propto f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{falt}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \\ &\propto f(\boldsymbol{y}_{falt}|\boldsymbol{y}_{obs}, \boldsymbol{\theta}) f(\boldsymbol{y}_{obs}|\boldsymbol{\theta}) f(\boldsymbol{\theta}). \end{aligned}$$

Em um algoritmo de Gibbs basta gerar da condicional completa de $\boldsymbol{y}_{falt}$ dada por $f(\boldsymbol{y}_{falt}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})$.

Exemplo

No modelo AR(1), suponha que tenha um dado faltante em $t^\star$ entre $(1, T)$.

A condicional completa de $y_{falt} \equiv y_{t^\star}$ é dada por

$$f(y_{t^\star}|y_{obs}, \boldsymbol{\theta}) \propto f(y_{t^\star}|y_{t^\star-1}, \boldsymbol{\theta}) f(y_{t^\star+1}|y_{t^\star}, \boldsymbol{\theta}).$$

Podemos mostrar que

$$(y_{t^\star}|y_{obs}, \boldsymbol{\theta}) \sim \mathcal{N}\left(\frac{\phi}{1+\phi^2}(y_{t^\star-1} + y_{t^\star+1}), \frac{\sigma^2}{1+\phi^2}.\right)$$

Algoritmo

1. Gerar $y_{t^\star}^{(k)}$ da distribução

$$(y_{t^\star}|y_{obs}, \boldsymbol{\theta}^{(k-1)}) \sim \mathcal{N}\left(\frac{\phi^{(k-1)}}{1+(\phi^{(k-1)})^2}(y_{t^\star-1} + y_{t^\star+1}), \frac{\sigma^{2(k-1)}}{1+(\phi^{(k-1)})^2}\right);$$

2. Completar os dados $\boldsymbol{y}^{(k)} = (\boldsymbol{y}_{1:(t^\star-1)}, y_{t^\star}^{(k)}, \boldsymbol{y}_{(t^\star+1):T})$; e
3. Passos usuais para gerar $(\sigma^2|\boldsymbol{y}^{(k)}, \phi^{(k-1)})$ e $(\phi|\boldsymbol{y}^{(k)}, \sigma^{2(k)})$.

Ver arquivo exemplo_71.r

Método da inicialização/reamostragem (*bootstrap*)

Um método de reamostrar um conjunto de dados.

Pode ser utilizado para inferência e também diagnóstico.

Aplica-se em muitos casos para avaliar a variância, o erro padrão e o viés de um estimador.

Evita cálculos analíticos. Entrentanto, é computacionalmente intensivo.

Na reamostragem paramétrica, consideramos modelos e reamostramos destes modelos utilizando as estimativas dos parâmetros.

A reamostragem não paramétrica é feita com base na distribuição empírica dos dados.

Consideraremos a amostra $(y_1, \ldots, y_n)$ provenientes de uma função (de densidade) de probabilidade f e função de distribuição acumulada F que dependem de um vetor de parâmetros $\boldsymbol{\theta}$.

Suponha que $T(\boldsymbol{y})$ seja a estatística de interesse para estimar $\boldsymbol{\theta}$ e $t(\boldsymbol{y})$ seu valor observado.

Reamostragem não paramétrica

A função de distruição empírica F_e é dada por

$$F_e(y) = \frac{1}{n} \sum_{i}^{n} \boldsymbol{I}_{[y_i, \infty)}(y),$$

que estima a função de distribuição F. Aqui, $\boldsymbol{I}_A(x)$ assume valor 1 se $x \in A$ e 0 caso contrário.

É possível escrever a estatística T em função de F_e. Isto nos dá uma expressão matemática para escrever T a partir de F.

Assim, $T(\boldsymbol{y}) = T(F_e(\boldsymbol{y}))$ e estimamos $\boldsymbol{\theta} = T(F_e(\boldsymbol{y}))$.

Exemplo

Seja $(y_1, \ldots, y_n)$ uma amostra aleatória de $F(y|\theta)$.

Tomemos $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$, a média e a variância amostral, respectivamente.

Estamos interessados em calcular a variância de $\overline{y}$ que na maioria dos casos depende de θ: $\text{Var}(\overline{y}|\theta)$. Em geral, teremos $\widehat{\text{Var}}(\overline{y}|\theta) = s^2/n$.

Seja $(y_1^\star, \ldots, y_n^\star)$ a reamostra de $(y_1, \ldots, y_n)$ com reposição.

Para cada reamostra de $m = 1, \ldots, M$, calcule $s^{2(m)}$.

A estimativa da variância via reamostragem é dada por $s_r^2 = \frac{1}{M}\sum_{m=1}^{M} s^{2(m)}$.

Ver arquivo exemplo_72.r

Reamostragem paramétrica

Na reamostragem paramétrica utilizamos a informação sobre a distribuição populacional dos dados.

Seja $(y_1, \ldots, y_n)$ uma amostra aleatória de $F(y|\boldsymbol{\theta})$.

Estimamos $\boldsymbol{\theta}$ com $\widehat{\boldsymbol{\theta}} = T(\boldsymbol{y})$ e substituímos em F para obter $\widehat{F}(y) = F(y|\widehat{\boldsymbol{\theta}})$.

Daí, reamostramos $(y_1^\star, \ldots, y_n^\star)$ de $F(y|\widehat{\boldsymbol{\theta}})$.

Para cada reamostra de $m = 1, \ldots, M$, calcule $s^{2(m)}$.

A estimativa da variância via reamostragem é dada por $s_r^2 = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} s^{2(m)}$.

Ver arquivo exemplo_73.r

Note que podemos facilmente estimar o viés do estimador do desvio padrão por

$$s_r - s, \quad \text{sendo} \quad s = \sqrt{s^2} \quad \text{e} \quad s_r = \frac{1}{M}\sum_{m=1}^{M} s^{(m)}.$$

Reamostragem: intervalo de confiança

Para cada reamostra, calculamos a estatística $T^{(m)}(\boldsymbol{y}^\star)$ e depois calcule

$$z^{(m)}(\boldsymbol{y}^\star) = \frac{T^{(m)}(\boldsymbol{y}^\star) - T(\boldsymbol{y})}{ep(T^{(m)}(\boldsymbol{y}^\star))}.$$

Estime o α-percentil de $z(\boldsymbol{y}^\star)$, t_α, pelo quantil amostral de $(z^{(1)}(\boldsymbol{y}^\star), \ldots, z^{(M)}(\boldsymbol{y}^\star))$.

Por fim, calcule o intervalo de confiança via reamostragem por

$$\text{IC}_{100(1-\alpha)\%}(\theta) = (T(\boldsymbol{y}) + t_{\alpha/2} \times \text{ep}(T(\boldsymbol{y})), T(\boldsymbol{y}) + t_{1-\alpha/2} \times \text{ep}(T(\boldsymbol{y}))).$$

Ver arquivos exemplo_72.r e exemplo_73.r

Reamostragem em modelos de regressão

O método de reamostragem pode ser utilizado para avaliar a precisão das estimativas estatísticas.

O método de reamostragem produz tipicamente uma aproximação para a distribuição da estatística de interesse que pode ser mais precisa que sua aproximação assintótica de primeira ordem.

Em modelos de regressão lineares temos

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \text{para} \quad i = 1, 2, \ldots, n,$$

em que ε_i são variáveis aleatórias independentes com média 0 e variância constante σ^2.

Sabemos que o estimador via mínimos quadrados ($\boldsymbol{X}$ de posto completo) é dado por

$$\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Neste exemplo, estamos interessados no erro padrão de β_k para $0 \leqslant k \leqslant p$.

- Estime β por $\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ e calcule o vetor de resíduos $\boldsymbol{e}$.
- Gere uma amostra $\boldsymbol{e}^\star$ de tamanho n, com reposição, de $\boldsymbol{e}$ e calcule $\boldsymbol{y}^\star = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}^\star$.
- Estime $\beta^\star$ de cada reamostra por $\boldsymbol{b}^\star = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^\star$.
- Por fim, calcule o erro padrão de β_k como o desvio padrão amostral das diversas estimativas $b_k^\star$ das reamostras.

Esse método de reamostragem dos resíduos funcionam bem quando o modelo é de fato o mais plausível como processo gerador dos dados.

Uma outra abordagem, que funciona inclusive em presença de heterocedasticidade, é reamostrar dos pares $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$.

Note que estamos fazendo estimação por mínimos quadrados e portanto não estamos assumindo nenhuma distribuição conhecida para o erro.

Ver arquivo exemplo_74.r

Método “deixe um fora” (*jackknife*)

Suponha que tenhamos uma amostra $\boldsymbol{y} = (y_1, \ldots, y_n)$ e um estimador $\widehat{\theta} = T(\boldsymbol{y})$.

O método “deixe um fora” (*jackknife*) considera as subamostras da forma

$$\boldsymbol{y}_i^\star = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \quad \text{para} \quad i = 1, \ldots, n.$$

e as estimativas

$$\widehat{\theta}_{(i)} = t(\boldsymbol{y}_i^\star) \quad \text{para} \quad i = 1, \ldots, n.$$

A estimativa do viés via este método é dada por

$$\widehat{\text{viés}} = (n-1)(\widehat{\theta}_{(.)} - \widehat{\theta}) \quad \text{em que} \quad \widehat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{(i)}.$$

A estimativa do erro padrão é dada por

$$\widehat{\text{ep}}(\widehat{\theta}) = \left[\frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}_{(i)} - \widehat{\theta}_{(.)})^2\right]^{1/2}.$$

Ver arquivo exemplo_75.r