Otimização numérica

Prof. Caio Azevedo

- Muitas vezes, na Estatística, temos por objetivo minimizar ou maximizar funções.
- Exemplos:
  - Obtenção de estimativas de máxima verossimilhança.
  - Obtenção de estimativas de mínimos quadrados.
  - Obtenção da moda da distribuição a posteriori.
  - Obtenção de intervalos de confiança de comprimento mínimo.
  - Minimzar critérios de perda
  - Maximizar critérios de otimalidade

- $X_i \overset{iid}{\sim} F_X(.,\boldsymbol{\theta})(f_X(x_i;\boldsymbol{\theta})), i = 1, ..., n$. Desejamos estimar $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^k$, com base na amostra aleatória.
- Verossimilhança: $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_x(x_i;\boldsymbol{\theta})$. Objetivo: maximizá-la.
- Logverossimilhança:$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_x(x_i;\boldsymbol{\theta})$.
- Caso univariado.
  - Função escore: $S(\theta) = \frac{d}{d\theta} l(\theta)$. Resolver $S(\widetilde{\theta}) = 0$.
  - Função Hessiana: $H(\theta) = \frac{d^2}{d\theta^2} l(\theta) = \frac{d}{d\theta} S(\theta)$.
  - Informação de Fisher: $I(\theta) = -\mathcal{E}_X(H(\theta))$ (em relação à X).

Caso Multiparamétrico

- Vetor escore:

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \begin{bmatrix} S_1(\boldsymbol{\theta}) \\ S_2(\boldsymbol{\theta}) \\ \vdots \\ S_k(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} l(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} l(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_k} l(\boldsymbol{\theta}) \end{bmatrix}$$

Caso Multiparamétrico

- Matriz Hessiana:

$$H(\boldsymbol{\theta}) = \begin{bmatrix} H_{11}(\boldsymbol{\theta}) & H_{12}(\boldsymbol{\theta}) & \dots & H_{1k}(\boldsymbol{\theta}) \\ H_{21}(\boldsymbol{\theta}) & H_{22}(\boldsymbol{\theta}) & \dots & H_{2k}(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1}(\boldsymbol{\theta}) & H_{k2}(\boldsymbol{\theta}) & \dots & H_{kk}(\boldsymbol{\theta}) \end{bmatrix}$$

Matriz Hessiana

- Matriz Hessiana:

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \theta_1 \partial \theta_k} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} l(\boldsymbol{\theta}) & \frac{\partial_2^2}{\partial \theta_2^2} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \theta_2 \partial \theta_k} l(\boldsymbol{\theta}) \\ \dots & \dots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_k \partial \theta_1} l(\boldsymbol{\theta}) & \frac{\partial^2 \theta_2}{\partial \theta_k \partial_1} l(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial \theta_k^2} l(\boldsymbol{\theta}) \end{bmatrix}$$

Informação de Fisher

- Informação de Fisher:

$$I(\boldsymbol{\theta}) = -\begin{bmatrix} \mathcal{E}\left(\frac{\partial^2}{\partial\theta_1^2} l(\boldsymbol{\theta})\right) & \mathcal{E}\left(\frac{\partial^2}{\partial\theta_1\partial\theta_2} l(\boldsymbol{\theta})\right) & \dots & \mathcal{E}\left(\frac{\partial^2}{\partial\theta_1\partial\theta_k} l(\boldsymbol{\theta})\right) \\ \mathcal{E}\left(\frac{\partial^2}{\partial\theta_2\partial\theta_1} l(\boldsymbol{\theta})\right) & \mathcal{E}\left(\frac{\partial^2\theta_2}{\partial\theta_2^2} l(\boldsymbol{\theta})\right) & \dots & \mathcal{E}\left(\frac{\partial^2}{\partial\theta_2\partial\theta_k} l(\boldsymbol{\theta})\right) \\ \dots & \dots & \ddots & \vdots \\ \mathcal{E}\left(\frac{\partial^2}{\partial\theta_k\partial\theta_1} l(\boldsymbol{\theta})\right) & \mathcal{E}\left(\frac{\partial^2\theta_2}{\partial\theta_k\partial\theta_1} l(\boldsymbol{\theta})\right) & \dots & \mathcal{E}\left(\frac{\partial^2}{\partial\theta_k^2} l(\boldsymbol{\theta})\right) \end{bmatrix}$$

Algoritmos de maximização

- Em geral, $S(\widetilde{\boldsymbol{\theta}}) = \mathbf{0}$, não apresentação solução explícita. Mesmo no caso univariado (modelos de regressão).
- Expansão em série de Taylor de $S(\theta)$ em torno de θ_0. Supõ-se que $|\theta - \theta_0| < \epsilon$. Assim

$$S(\theta) = S(\theta_0) + (\theta - \theta_0)H(\theta_0)$$

$$\Rightarrow \theta = \theta_0 - H(\theta_0)^{-1}S(\theta_0)$$

Newton-Raphson

- Inicie com uma aproximação razoável $\boldsymbol{\theta}^{(0)}$ e faça, t=0,1,2,3,..

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - H(\boldsymbol{\theta}^{(t)})^{-1} S(\boldsymbol{\theta}^{(t)})$$

  até que algum critério de convergência seja obtido.

- $|\theta^{(t+1)} - \theta^{()}| < \epsilon$ ou $|l(\theta^{(t+1)}) - l(\theta^{(t)})| < \delta$.

Características

- Funciona bem para os casos uniparamétricos e/ou com verossimilhanças regulares.
- Sensível à escolha dos valores iniciais.
- Necessita da obtenção analítica da matriz Hessiana.

Alternativa: A matriz Hessiana pode ser substituída pela informação de Fisher.

Escore de Fisher

- Inicie com uma aproximação razoável $\boldsymbol{\theta}^{(0)}$ e faça, t=0,1,2,3,..

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + I(\boldsymbol{\theta}^{(t)})^{-1} S(\boldsymbol{\theta}^{(t)})$$

  até que algum critério de convergência seja obtido.

- $|\theta^{(t+1)} - \theta^{()}| < \epsilon$ ou $|I(\theta^{(t+1)}) - I(\theta^{(t)})| < \delta$.

Carcterísticas

- Funciona bem para os casos uniparétricos e/ou com verossimilhanças regulares.
- Sensível à escolha dos valores iniciais.
- Necessita da obtenção analítica da matriz Hessiana.
- Mais estável do que ao algoritmo de NR.
- Mais simples de ser implementado do que o algoritmo de NR desde que o cálculo da Informação de Fisher não seja custoso.

- Seja $X_1, ..., X_n$ uma amostra aleatória de $X \sim gama(r, \theta)$, θ conhecido
- Verossimilhança

$$L(r) = e^{-\sum_{i=1}^{n} \frac{x_i}{\theta}} \prod_{i=1}^{n} x_i^{r-1} \frac{1}{(\Gamma(r))^n \theta^{nr}}$$

- Log-verossimilhança

$$l(r) = -n \ln(\Gamma(r)) - nr \ln(\theta) - \sum_{i=1}^{n} \frac{x_i}{\theta} + (r-1) \sum_{i=1}^{n} \ln(x_i)$$

- Função escore

$$\begin{aligned} S(r) &= -n\frac{\Gamma'(r)}{\Gamma(r)} - n\ln\theta + \sum_{i=1}^{n}\ln(x_i) \\ &= -n\Psi(r) - n\ln\theta + \sum_{i=1}^{n}\ln x_i \end{aligned} \quad (1)$$

- Função Hessiana

$$H(r) = -\frac{n}{(\Gamma(r))^2}\left[\Gamma''(r)\Gamma'(r) - (\Gamma'(r))^2\right] = -n\Psi'(r)$$

em que $\Psi(r)$ e $\Psi'(r)$ são, respectivamente, a função digama e a função trigama.

- Informação de Fisher

$$I(\theta) = \frac{n}{(\Gamma(r))^2}\left[\Gamma''(r)\Gamma'(r) - (\Gamma'(r))^2\right] = n\Psi'(r)$$

- A equação $S(\widetilde{r}) = 0$ não possui solução explícita.
- Solução: utilização de métodos numéricos para obtenção de raízes de equações (não-lineares).
- Algoritmo de Newton-Raphson.
- Versão estocástica : Algoritmo escore de Fisher.
- Suporte teórico: expansão em séries de Taylor da função Escore.

- É um método iterativo.
- Inicializa-se com um "chute" inicial (de preferências um valor próximo do verdadeiro valor do parâmetros ou dos parâmetros).
- Com o chute inicial, gera-se uma nova estimativa para o parâmetro.
- Verifica-se se o critério de parada foi satisfeito.
- Se sim, termina-se o processo, caso contrário, gera-se um novo valor para o parâmetro com base na estimativa anterior.
- Repete-se os dois passos anteriores, até que o critério de parada seja alcançado.

- Seja $\widetilde{\boldsymbol{\theta}}^{(t)}$ uma estimativa para $\boldsymbol{\theta}$ obtido na iteração t.
- Obtem-se uma estimativa atualizada de $\boldsymbol{\theta}$, digamos $\widetilde{\boldsymbol{\theta}}^{(t+1)}$, através de

$$\widetilde{\boldsymbol{\theta}}^{(t+1)} = \widetilde{\boldsymbol{\theta}}^{(t)} + \mathbf{I}\left(\widetilde{\boldsymbol{\theta}}^{(t)}\right)^{-1} \mathbf{S}\left(\widetilde{\boldsymbol{\theta}}^{(t)}\right)$$

- Critério de parada: por exemplo $||\widetilde{\boldsymbol{\theta}}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}|| < \epsilon, \epsilon > 0$

- Distribuição assintótica de $\widehat{r}$ (emv de r). Para n suficientemente grande, temos que

$$\widehat{r} \approx N\left(r, \frac{1}{n\Psi'(r)}\right)$$

- Erro-padrão assintótico de $\widehat{r}$, $EP_A(\widehat{r}) = \sqrt{\frac{1}{n\Psi'(r)}}$.

Cont.: ambos os parâmetros desconhecidos

- Vetor escore:

$$S(r,\theta) = \begin{bmatrix} -n\Psi(r) - n\ln\theta + \sum_{i=1}^{n} \ln x_i \\ \frac{n\bar{x}}{\theta^2} - \frac{nr}{\theta} \end{bmatrix}$$

- Matriz Hessiana:

$$\mathbf{H}(r,\theta) = \begin{bmatrix} -n\Psi'(r) & -\frac{n}{\theta} \\ -\frac{n}{\theta} & \frac{nr}{\theta^2} - \frac{2n\bar{x}}{\theta^3} \end{bmatrix}$$

- Informação de Fisher:

$$\mathbf{I}(r,\theta) = \begin{bmatrix} n\Psi'(r) & \frac{n}{\theta} \\ \frac{n}{\theta} & \frac{nr}{\theta^2} \end{bmatrix}$$

Comentários

- O cálculo analítico da matriz Hessiana e/ou da Informação de Fisher pode ser custoso.
- Os algoritmos de NR/SF são sensíveis à escolha de valores iniciais e à comportamentos não regulares da verossimilhança.
- Os algoritmos Quasi-Newton utilizam idéias semelhantes ao algoritmos anteriories. Contudo:
  - Trabalham com aproximações numéricas das matrizes Hessinas.
  - Atualizam as estimativas de modo diferente (tentanto evitar máximos/mínimos locais).

Algoritmo **B**royden**F**letcher**G**oldfarb**S**hanno (BFGS)

- Seja H_A uma aproximação analítica para a matrix Hessiana (pode ser calculada analiticamente).
- Inicie com uma aproximação razoável $\boldsymbol{\theta}^{(0)}$ e $H_A(\boldsymbol{\theta}^{(0)})$. Se $S(\boldsymbol{\theta}^{(0)}) = \mathbf{0}$ pare, caso contrário faça para t=0,1,2,3,...
- $d^{(t)} = -H_A(\boldsymbol{\theta}^{(t)})^{-1} S(\boldsymbol{\theta}^{(t)})$
- Para $\alpha^{(t)} > 0$ calcule $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \alpha^{(t)} d^{(t)}$
- Repetir até que $S(\boldsymbol{\theta}^{(t+1)})$.