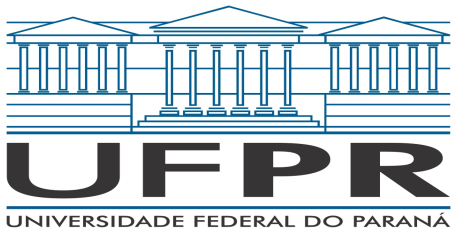
Máxima Verossimilhança

Prof. Wagner H. Bonat

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação



Sumário

Notação e Definições

- O vetor $(n \times 1)$ de variáveis aleatórias (va) é denotado por $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.
- O vetor $(n \times 1)$ de realizações de uma va é denotado por $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$.
- Denote $f(\boldsymbol{Y}; \boldsymbol{\theta})$ a função de probabilidade (fp) caso discreto ou função de densidade probabilidade (fdp) do vetor aleatório $\boldsymbol{Y}$.
- **Parâmetro ou vetor de parâmetros** Vetor de características numéricas de uma população.
- Notação: $\boldsymbol{\theta}$ vetor $(p \times 1)$ de parâmetros.
- **Espaço paramétrico** é o conjunto de todas as possíveis combinações entre todos os valores para todos os diferentes parâmetros envolvidos em uma fp ou fdp. Notação Θ.
- **Suporte** é conjunto de valores realizáveis de uma va. Notação Ω.
- Exemplos: Binomial, Poisson e normal.

Definições

- Uma **estatística** é uma variável aleatória $T = t(\boldsymbol{Y})$, onde a função $t(\cdot)$ não depende de θ.
- Uma **estatística** T é um **estimador** para θ se o valor realizado $t = t(\boldsymbol{y})$ é usado como uma **estimativa** para o valor de θ.
- A distribuição de probabilidade de T é chamada de **distribuição amostral** do estimador $t(\boldsymbol{Y})$.
- O **viés** de um estimador T é a quantidade

$$B(T) = E(T - \theta).$$

  O estimador T é dito não viciado para θ se $B(T) = 0$, tal que $E(T) = \theta$.
- O estimador T é assintóticamente não viciado para θ se $E(T) \to \theta$ quando $n \to \infty$.

Definições

- A **eficiência relativa** entre dois estimadores T_1 e T_2 é a razão $er = \frac{V(T_1)}{V(T_2)}$ em que $V(\cdot)$ denota a variância.
- O **erro quadrático médio** de um estimador T é a quantidade

$$EQM(T) = E((T-\theta)^2) = V(T) + B(T)^2.$$

- Um estimador T é **médio quadrático consistente** para θ se o $EQM(T) \to 0$ quando $n \to \infty$.
- O estimador T é **consistente em probabilidade** se $\forall \epsilon > 0$, $P(|T-\theta| > \epsilon) \to 0$, quando $n \to \infty$.

Sumário

Função de verossimilhança

- Sejam dados $\boldsymbol{y}$ uma realização de um vetor aleatório $\boldsymbol{Y}$ com fp ou fdp $f(\boldsymbol{Y}, \boldsymbol{\theta})$. A **função de verossimilhança** $L(\boldsymbol{\theta}) : \Theta \to [0, \infty]$ para $\boldsymbol{\theta}$ é a função aleatória

$$L(\boldsymbol{\theta}) \equiv f(\boldsymbol{Y}, \boldsymbol{\theta})$$

onde $f(y_1, \ldots, y_n | \boldsymbol{\theta})$ é a função de distribuição conjunta de $\boldsymbol{Y}$.

1. Caso discreto não há ambiguidade então

$$L(\boldsymbol{\theta}) \equiv P_{\boldsymbol{\theta}}[Y = y].$$

2. Caso contínuo em geral as observações são medidas com algum grau de precisão em um intervalo $(y_{iI} \leq y_i \leq y_{iS})$. Neste caso a verossimilhança é dada por

$$L(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}[y_{1I} \leq y_1 \leq y_{1S}, y_{2I} \leq y_2 \leq y_{2S}, \ldots, y_{nI} \leq y_n \leq y_{nS}].$$

Função de verossimilhança

- Suponha que as observações são independentes e medidas com o mesmo grau de precisão.
- Assim, cada dado é medido em um intervalo $(y_i - \delta/2 \leq Y_i \leq y_i + \delta/2)$.
- Com estas suposições a verossimilhança pode ser escrita como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P_{\boldsymbol{\theta}}[y_i - \delta/2 \leq Y_i \leq y_i + \delta/2]$$
$$= \prod_{i=1}^{n} \int_{y_i - \delta/2}^{y_i + \delta/2} f(y_i, \theta) d(y_i).$$

Função de verossimilhança

- Se o grau de precisão é alto (δ é pequeno) em relação a variabilidade dos dados, a expressão se reduz a

$$L(\boldsymbol{\theta}) \approx \left(\prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \right) \delta^n.$$

- Finalmente, se δ não depende de $\boldsymbol{\theta}$, temos

$$L(\boldsymbol{\theta}) \approx \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}),$$

- Para enfatizar que a verossimilhança é avaliada nas observações usamos a notação $L(\boldsymbol{\theta}|\boldsymbol{y})$.

Função de Log-Verossimilhança

- A função de log-verossimilhança é a função estocástica $l(\boldsymbol{\theta}) : \Theta \to \Re$ definida por
$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \log\left(L(\boldsymbol{\theta}|\boldsymbol{y})\right).$$
- No caso iid, tem-se
$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log\left(L(\boldsymbol{\theta}|y_i)\right).$$
- $l(\boldsymbol{\theta}|\boldsymbol{y}) = -\infty$ quando $L(\boldsymbol{\theta}) = 0$, mas isso ocorre quando $f(y_1, \ldots, y_n|\boldsymbol{\theta}) = 0$ que tem probabilidade de ocorrência igual a zero.

Vetor escore

- O vetor escore $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) : \Theta \to \Re^p$ é um vetor aleatório $p \times 1$ definido por

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \theta_p} \end{pmatrix}.$$

- Notação popular em termos de gradiente

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{y}).$$

- Notação j-ésimo componente de $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y})$ por $\boldsymbol{U}_j(\boldsymbol{\theta}|\boldsymbol{y})$.

Matriz de informação esperada (Fisher Information Matrix)

- A matriz $p \times p$ definida por

$$\begin{aligned} \mathbf{I}_E(\boldsymbol{\theta}) &= \mathbf{Var}(\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y})) \\ &= \mathrm{E}(\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y})\boldsymbol{U}^{\top}(\boldsymbol{\theta}|\boldsymbol{Y})). \end{aligned}$$

  é chamada de matriz de informação esperada.
- As entradas j e k são expressadas por

$$\begin{aligned} \mathbf{I}_{Ejk}(\boldsymbol{\theta}) &= \mathrm{Cov}(\boldsymbol{U}_j(\boldsymbol{\theta}|\boldsymbol{Y}), \boldsymbol{U}_k(\boldsymbol{\theta}|\boldsymbol{Y})) \\ &= \mathrm{E}(\boldsymbol{U}_j(\boldsymbol{\theta}|\boldsymbol{Y})\boldsymbol{U}_k(\boldsymbol{\theta}|\boldsymbol{Y}))). \end{aligned}$$

Matriz de informação observada

- A matriz $p \times p$ definida por

$$\mathrm{I}_O(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}$$

  é chamada de informação observada.
- As entradas j e k da matriz de informação observada é dada por

$$\mathrm{I}_{Ojk}(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial\theta_j\partial\theta_k}.$$

- Primeira igualdade de Bartlett $E(\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{Y})) = \mathbf{0}$.
- Segunda igualdade de Bartlett

$$\mathrm{I}_E(\boldsymbol{\theta}) = \mathrm{E}(\mathrm{I}_O(\boldsymbol{\theta})).$$

Estimador de máxima verossimilhança (EMV)

- Estimativa de máxima verossimilhança: Seja $L(\boldsymbol{\theta}, \boldsymbol{y})$ a função de verossimilhança. O valor $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$ é a estimativa de máxima verossimilhança para $\boldsymbol{\theta}$ se $L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta}$.
- Estimador de máxima verossimilhança: Se $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ é a estimativa de máxima verossimilhança, então $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ é o estimador de máxima verossimilhança (EMV).
- Em geral, $\hat{\boldsymbol{\theta}}$ satisfaz a equação de verossimilhança

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \mathbf{0}.$$

- Um sistema com p equações e p incógnitas

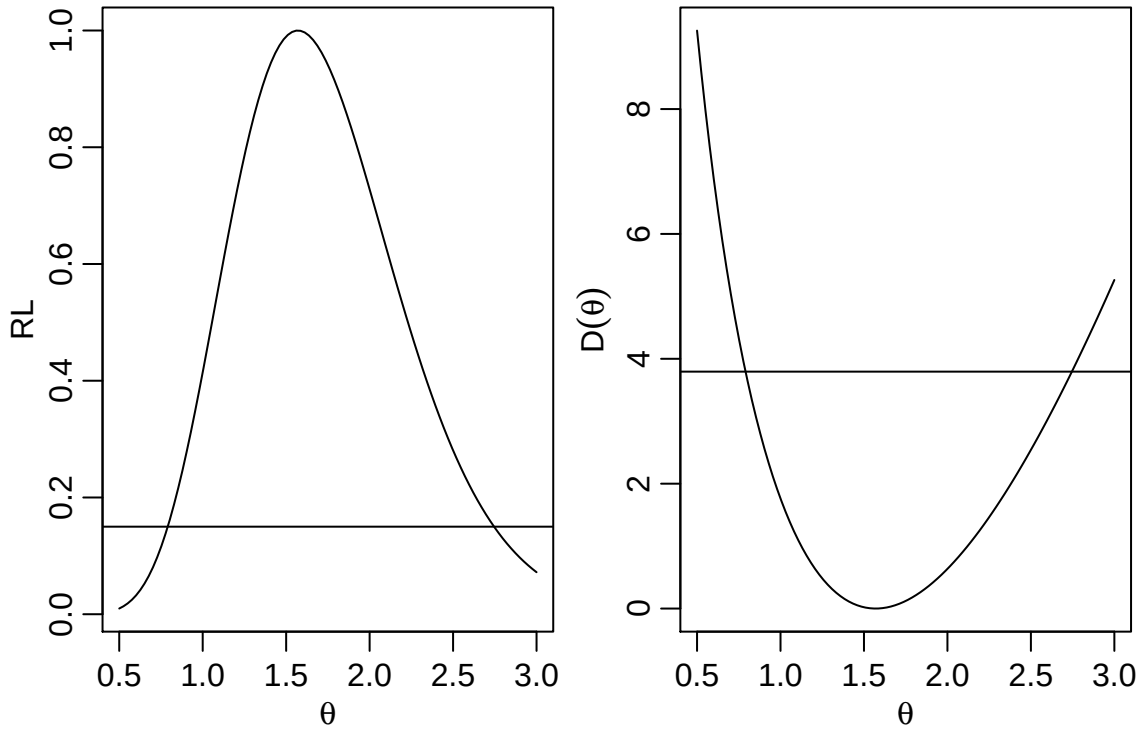
$$\begin{pmatrix} U_1(\boldsymbol{\theta}|\boldsymbol{y}) = 0 \\ \vdots \\ U_p(\boldsymbol{\theta}|\boldsymbol{y}) = 0 \end{pmatrix}$$

Intervalos de confiança

- Um intervalo de verossimilhança para θ é um intervalo da forma $\theta : L(\theta) \geq rL(\hat{\theta})$ ou equivalentemente, $\theta : D(\theta) \leq c^*$, com $D(\theta) = -2[l(\theta) - l(\hat{\theta})]$ e $c^* = -2\log(r)$.
- Para o caso multiparâmetros a ideia se mantêm, mas trocamos o intervalo por uma região de confiança.
- Nesta definição r precisa ser uma valor entre 0 e 1, logo $c^* > 0$ para intervalos não-vazios.

Estratégias para construir o intervalo de confiança

- Verossimilhança relativa $\frac{L(\theta)}{L(\hat{\theta})} \geq r$.
- Deviance $D(\theta) = -2[l(\theta) - l(\hat{\theta})] \leq -2\log(r)$.
- Após definir o valor $c^* = -2\log(r)$, é necessário encontrar as raízes da função de verossimilhança relativa ou da *deviance*.



Estratégias para construir o intervalo de confiança

- De forma geral métodos numéricos serão necessários.
- Alternativa: Expansão em séries de Taylor (segunda ordem) para $l(\theta)$ em torno de $\hat{\theta}$

$$\begin{aligned} D(\theta) &= -2[l(\theta) - l(\hat{\theta})] \\ &\approx 2\left\{l(\hat{\theta}) - [l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})]\right\}. \end{aligned}$$

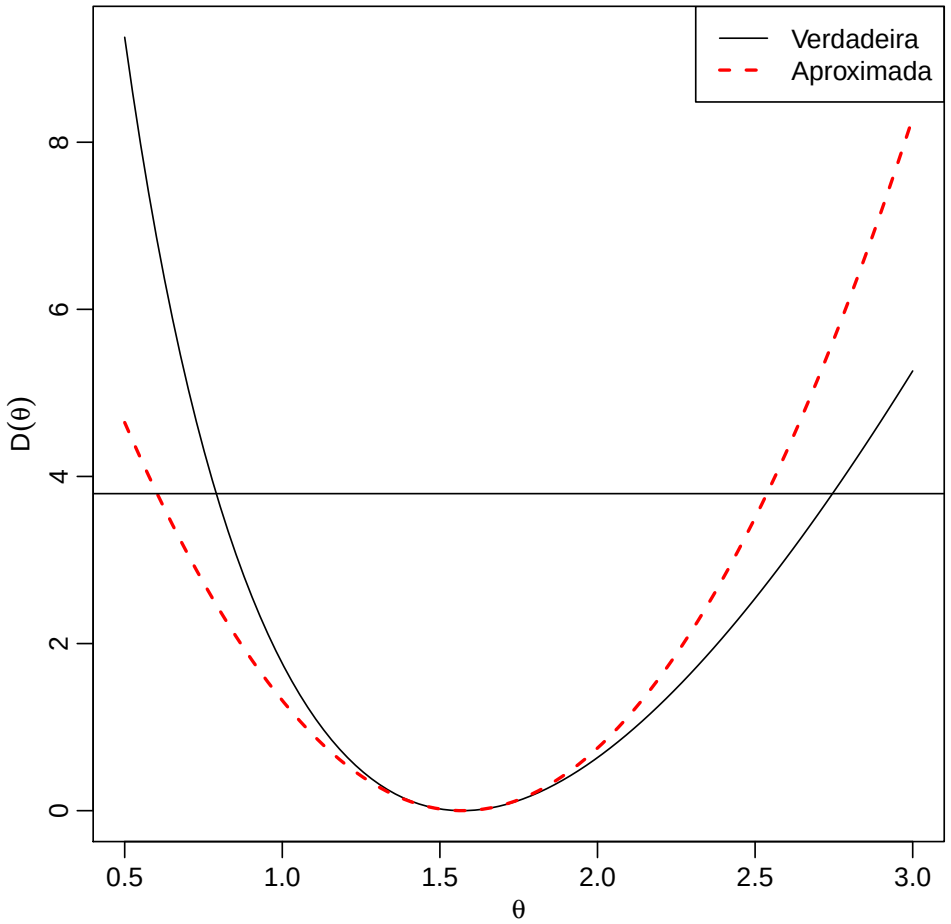
- Eliminando termos, temos

$$D(\theta) \approx -(\theta - \hat{\theta})^2 l''(\hat{\theta}) \leq c^*.$$

- Resolvendo em θ chegamos a intervalos da forma.

$$\hat{\theta} \pm \sqrt{\frac{c^*}{-l''(\hat{\theta})}}.$$

Estratégias para construir o intervalo de confiança



- Ainda precisamos definir r ou c^*.

Condições de regularidade

- O parâmetro θ é **identificável**. Isso significa que se $f(\theta_1|\boldsymbol{y}) = f(\theta_2|\boldsymbol{y})$ para quase todos $y \in \Re$, então $\theta_1 = \theta_2$.
- O suporte de $f(\theta|\boldsymbol{y})$ é o mesmo para todo $\theta \in \Re$.
- O verdadeiro valor do parâmetro θ_0 pertence ao interior de Θ.
- $f(\boldsymbol{\theta}|\boldsymbol{y})$ é duas vezes continuamente diferenciável com relação aos componentes de $\boldsymbol{\theta}$ para quase todo $\boldsymbol{y} \in \Re$.
- $\frac{\partial}{\partial\theta}$ e $\int$ (caso contínuo), ou $\frac{\partial}{\partial\theta}$ e $\sum$ (caso discreto) podem ser intercambiados.

Propriedades do Estimador de máxima verossimilhança

- Sendo $\boldsymbol{\theta}_0$ o verdadeiro valor do vetor de parâmetros $\boldsymbol{\theta}$. Então
- Consistência: $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$, ou seja,

$$P\left(\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| > \epsilon\right) \to 0, \quad \text{quando} \quad n \to \infty.$$

- Normalidade assintótica:

$$\hat{\boldsymbol{\theta}} \xrightarrow{D} N_p(\boldsymbol{\theta}, \mathrm{I}_E^{-1}(\boldsymbol{\theta})), \quad \text{quando} \quad n \to \infty.$$

- Qualquer termo assintoticamente equivalente a $I_E(\boldsymbol{\theta})$ pode ser usado.

$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, I_E^{-1}(\hat{\boldsymbol{\theta}}))$$
$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, I_O^{-1}(\boldsymbol{\theta}))$$
$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, I_O^{-1}(\hat{\boldsymbol{\theta}})).$$

Distribuição assintótica da deviance

- Para um problema regular de estimação, no limite com $n \to \infty$, se $\boldsymbol{\theta}$ é o verdadeiro valor do parâmetro, então

$$D(\boldsymbol{\theta}) = -2[l(\boldsymbol{\theta}) - l(\hat{\boldsymbol{\theta}})] \sim \chi^2_p$$

ou seja, a função deviance segue uma distribuição Qui-Quadrado com d graus de liberdade, onde p é a dimensão do vetor $\boldsymbol{\theta}$.

Resumo dos resultados

- O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ é assintóticamente não-viciado, isto é, $E(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\theta}$.
- Assintóticamente $V(\hat{\boldsymbol{\theta}}) \to I_E^{-1}(\boldsymbol{\theta})$, o qual por uma versão multivariada do limite de Cramér-Rao é o melhor possível, mostrando que o EMV é eficiente para o vetor $\boldsymbol{\theta}$.
- Denote $J = I_E^{-1}(\boldsymbol{\theta})$, então $V(\hat{\boldsymbol{\theta}}) = J$, sendo que, J é uma matriz simétrica e definida positiva, com elementos $J_{ij} = Cov(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j)$ então J_{ii} é a variância de $\hat{\boldsymbol{\theta}}_i$.
- Denota-se $J_{ii}^{\frac{1}{2}}$ de desvio padrão de $\hat{\boldsymbol{\theta}}_i$.
- Podemos construir intervalos de $100(1-\alpha)\%$ de confiança para θ_i na forma $\hat{\theta}_i \pm z_{\frac{\alpha}{2}} J_{ii}^{\frac{1}{2}}$. Intervalos desta forma serão denominados, intervalos de Wald ou baseados em aproximação quadrática da verossimilhança.

Regiões de confiança

- Para regiões de confiança baseados na *deviance* considera-se $\boldsymbol{\theta} \in \Theta : D(\boldsymbol{\theta}) \leq c^*$, para algum valor c^* a ser especificado. Pode-se escolher c^* baseado em justificativas assintóticas de que $D(\boldsymbol{\theta}) \sim \chi^2_p$ é uma escolha razoável para $c^* = c_\alpha$ com $P(\chi^2_p \geq c_\alpha) = \alpha$, por exemplo se $\alpha = 0.05$ e $p = 1$, então $c_\alpha = 3.84$. Isto gera uma região de $100(1-\alpha)\%$ de confiança. Estes intervalos serão denominados de intervalos *deviance*.
- Note que intervalos *deviance* e baseados em verossimilhança são de certa forma equivalentes. A diferença é a justificativa para a forma da região de confiança.
- Discutir os tutoriais I e II.

Método delta

- Considere obter um intervalo de confiança para $\phi = g(\boldsymbol{\theta})$ por invariância temos que $\hat{\phi} = g(\hat{\boldsymbol{\theta}})$ e a variância de $\hat{\phi}$ é dada por

$$V(\hat{\phi}) = V(g(\hat{\boldsymbol{\theta}})) = \nabla g(\hat{\boldsymbol{\theta}})^\top I_E(\hat{\boldsymbol{\theta}})^{-1} \nabla g(\hat{\boldsymbol{\theta}})$$

  onde

$$\nabla g(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial g(\hat{\boldsymbol{\theta}})}{\partial \theta_1}, \ldots, \frac{\partial g(\hat{\boldsymbol{\theta}})}{\partial \theta_d} \right)^\top .$$

- De imediato, temos

$$\hat{\phi} \sim N(\phi, \nabla g(\boldsymbol{\theta})^\top I_E(\boldsymbol{\theta})^{-1} \nabla g(\boldsymbol{\theta})).$$

- Ver tutorial III.

Verossimilhança perfilhada

- A verossimilhança perfilhada de $\boldsymbol{\theta}$ é definida por

$$\tilde{L}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}).$$

- A forma apresentada sugere um procedimento de maximização em duas etapas. A primeira consiste em obter $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}$ que maximiza $l(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \log L(\boldsymbol{\theta}, \boldsymbol{\lambda})$ com respeito a $\boldsymbol{\lambda}$ supondo $\boldsymbol{\theta}$ fixo.
- A seguir maximiza-se $\tilde{l}(\boldsymbol{\theta})$. Assim, uma região ou intervalo de confiança para $\boldsymbol{\theta}$ pode ser obtida usando que

$$\tilde{D}(\boldsymbol{\theta}) = -2[\tilde{l}(\boldsymbol{\theta}) - \tilde{l}(\hat{\boldsymbol{\theta}})] \sim \chi^2_p$$

onde p é a dimensão de $\boldsymbol{\theta}$.

Parâmetros ortogonais

- Considere um modelo estatístico parametrizado por $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\top$. No caso da matriz de informação de Fisher ser bloco diagonal

$$\mathrm{I}_E(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}_{E_1}(\boldsymbol{\theta}) & 0 \\ 0 & \mathrm{I}_{E_2}(\boldsymbol{\theta}) \end{pmatrix},$$

  os vetores de parâmetros $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ são ditos **ortogonais**.
- A distribuição assintótica de $\boldsymbol{\theta}_1$ é a mesma se $\boldsymbol{\theta}_2$ é considerado conhecido ou desconhecido.
- Definição similar pode ser feita usando a matriz de informação observada.
- Ver tutorial II.