Wyklad6

10 listopada 2012
08:44

1.3 Wieloparametrowe ograniczenie C-R

$p_{\vec{\theta}}(x)$ $\qquad \vec{\theta} = (\theta_1, \ldots, \theta_k) =: \theta$

$\vec{x} = (x_1, \ldots, x_N) =: x$

$$\boxed{C_{\tilde{\theta}} \geq \mathbb{F}^{-1}} \qquad \{ C_\theta - \mathbb{F}_\theta^{-1} \geq 0$$

macierz kowariancji $\qquad$ macierz Fishera

$$(C_{\tilde{\theta}})_{ij} = \int dx\, p_\theta(x) \left(\tilde{\theta}_i(x) - \theta_i \right) \left(\tilde{\theta}_j(x) - \theta_j \right)$$

gdzie $(\mathbb{F})_{ij} = \int dx \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial \theta_i} \frac{\partial p_\theta(x)}{\partial \theta_j}$ - Macierz Fishera (symetryczna dodatnio określona)

$$\{ \mathbb{F}_{ij} = \left\langle - \frac{\partial^2 \log p_\theta(x)}{\partial \theta_i \partial \theta_j} \right\rangle$$

W szczególności:

$$\Delta^2 \theta_i \geq (\mathbb{F}^{-1})_{ii}$$

{ Ma macierz dodatnio określonej wyrazy diagonalne nieujemne

$\mathbb{F}$ - macierz symetryczna, $\mathbb{F} \geq 0$

Dowód:

$$\int dx\, \tilde{\theta}_i(x)\, p_{\vec{\theta}}(x) = \theta_i$$

$$\int dx\, \tilde{\theta}_i(x) \frac{\partial p_{\vec{\theta}}(x)}{\partial \theta_j} = \delta_{ij} \quad \text{- lokalnie nieobciążoność}$$

$$\int dx\, \tilde{\theta}(x) \cdot \left(\nabla_\theta p_{\vec{\theta}}(x) \right)^T = \mathbb{1}$$

$$\int dx\, \nabla p_\theta(x) = 0 \quad \text{regularność}$$

{ Niech a, b dowolne wektory długości k

$$\int dx\, a^T (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^T a\, p_\theta(x) \cdot$$

$$\cdot \int \frac{1}{p_\theta(x)}\, b^T (\nabla p_\theta(x))(\nabla p_\theta(x))^T b \overset{C-S}{\geq}$$

$$\left(\int dx\ b^T \nabla p_\theta(x) (\tilde\theta - \theta) a\right)^2$$

$$a^T C_{\tilde\theta}\, a \cdot b^T F b \geq (b^T a)^2$$

Ma być spełnione dla dowolnych a i b

Np. biorąc $a = b = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix}$ - i-te miejsce

$$\Delta^2 \tilde\theta_{ii} \cdot F_{ii} \geq 1 \qquad \Delta^2 \tilde\theta_{ii} \geq \frac{1}{F_{ii}}$$

ale możemy lepsze ograniczenie

Weźmy $b = F^{-1} a$

$$a^T C_{\tilde\theta}\, a\ a^T F^{-1} a \geq (a^T F^{-1} a)^2$$

$$a^T (C_{\tilde\theta} - F^{-1})\, a \geq 0$$

$$C_{\tilde\theta} \geq F^{-1}$$

$$\Delta^2 \tilde\theta_{ii} \geq (F^{-1})_{ii} \geq (F_{ii})^{-1}$$

mocniejsze

$$1 = e_i^T \sqrt{F} \sqrt{F^{-1}} e_i \leq e_i^T F e_i \quad e_i^T F^{-1} e_i$$

$$(F^{-1})_{ii} \geq \frac{1}{F_{ii}}$$

Wysycenie:

$$\vec\nabla \ln p_\theta(x) = F(\vec g(x) - \vec\theta)$$

1.4 Modele liniowe

$$\vec x = H \cdot \vec\theta + \vec w \qquad \vec x = [x_1, \dots, x_N]$$

$$\vec\theta = [\theta_1, \dots \theta_K]$$

$$w_i \sim N(0, \sigma^2) \qquad \vec w = [w_1, \dots w_N]$$

Jeśli $N > K$ i H macierz rzędu K to

$\vec x \sim N(H\vec\theta, \sigma^2 \cdot \mathbb{1})$ – wielowymiarowy gauss o średniej $H\vec\theta$ i macierzy kowariancji $\sigma^2\mathbb{1}$

$$p(\vec x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}(\vec x - H\vec\theta)^T(\vec x - H\vec\theta)}$$

Warunek na wysycenie:

$$\vec\nabla \log p(\vec x) = -\frac{1}{2\sigma^2} \cdot \vec\nabla_\theta \left(\vec x^T x + \vec\theta^T H^T H \theta \underbrace{- \vec\theta^T H^T \vec x - x^T H \vec\theta}_{-2x^T H\vec\theta} \right) =$$

$$= -\frac{1}{2\sigma^2} \cdot \left(-2 H^T \vec x + 2 H^T H \theta \right) = \frac{1}{\sigma^2}\left(H^T \vec x - H^T H \vec\theta \right)$$

$$\begin{cases} \vec\nabla_\theta (a^T\theta) = a & \left\{ \frac{\partial}{\partial\theta_j} \sum_i a_i \theta_i = a_j \right. \\ \vec\nabla_\theta (\theta^T A \theta) = 2A\theta & \left\{ \frac{\partial}{\partial\theta_k} \left(\sum_{ij} \theta_i A_{ij} \theta_j = \sum_i \theta_i A_{ik} + {}_k A_{ki}\theta_i = 2A_{ki}\theta_i \right. \right. \\ \quad \text{macierz sym} & \end{cases}$$

Jeśli $H^T H$ odwracalne to:

$$\vec\nabla \log p(x) = \frac{1}{\sigma^2}(H^T H) \Big(\underbrace{(H^T H)^{-1} H^T \vec x}_{\tilde{\vec\theta}} - \vec\theta \Big)$$

estymator $\tilde{\vec\theta} = (H^T H)^{-1} H^T \vec x$

Macierz Fishera: $F = \frac{H^T H}{\sigma^2}$

– dopasowanie krzywych.

1.3 Estymacja funkcji od wielu parametrów

$g(\vec{\theta})$ } g również może być wieloparametrowe, ale dla uproszczenia niech skalarna

$$\Delta^2 \tilde{g} \geq \vec{\nabla} g \cdot F^{-1} (\vec{\nabla} g)^T$$

2. Estymacja największej wiarygodności (max-likelihood)

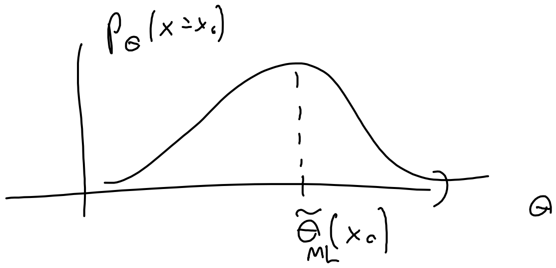
Co zrobić kiedy warunku wysycenia C-R nie spełniamy i nie potrafimy podać optymalnego estymatora.

2.1 Estymator największej wiarygodności

$p_\theta(x)$

jeśli mierzymy x_0 to jako estymator przyjmujemy taką wartość θ dla której $p_\theta(x)$ maksymalne

$p_\theta(x = x_0)$ – likelihood function



Intuicja: wartość parametru, dla którego obserwowane zdarzenie najbardziej prawdopodobne

Uwaga

Zazwyczaj wygodniej szukać max $\log p_\theta(x=x_0)$ a to do tego samego punktu bo log monotoniczny

Przykład

$$x_i \sim N(\theta, \sigma^2)$$

$$\frac{d}{d\theta} \log p_\theta(x) = 0 \qquad \text{- szukamy max}$$

$$\frac{1}{\sigma^2}\sum_i (x_i - \theta) = 0 \quad \Rightarrow \quad \tilde{\theta}_{ML}(x) = \frac{1}{N}\sum_i x_i$$

Ogólna własność - jeśli istnieje estymator efektywny - wysycający C-R to zawsze to ten sam, który znajdziemy max-likelihood

Pamiętamy że warunek wysycania to:

$$\frac{d}{d\theta} \log p_\theta(x) = \lambda(\theta)\left(f(x) - \theta \right)$$

$f(x)$ - efektywny estymator

czyli θ dla którego minimum to $\tilde{\theta}_{ML} = f(x)$

W szczególności dla modeli liniowych

2.2. Asymptotyczna efektywność Max-likelihood

W ogólności nie ma gwarancji, że max-likelihood wysyci C-R ani, że jest najlepszym estymatorem

ale ... jeśli dużo niezależnych obserwacji to tak

Twierdzenie

$$p_\theta^{(N)}(x^{(N)}) = p_\theta(x_1) \cdot \ldots \cdot p_\theta(x_N)$$

$x^{(N)} = [x_1, \ldots, x_N]$

x_i może reprezentować wiele danych

N powtórzeń eksperymentu

i wtedy w granicy $N \to \infty$ estymator wtedy w granicy $N \to \infty$, estymator ML będzie miał rozkład.

$$\tilde{\Theta}_{ML} \overset{N\to\infty}{\sim} \mathcal{N}\left(\Theta, F^{(N)^{-1}}\right), \quad F^{(N)} = N \cdot F$$

czyli wysyca C-R. i jest nieobciążony

<u>Lemat:</u>

Niech $p_1(x)$, $p_2(x)$ - rozkłady prawdop.

$$\int p_1(x) \log \frac{p_1(x)}{p_2(x)} \geq 0$$

$$\parallel$$

$$D(p_1 | p_2) \quad - \text{ entropia względna}$$

<u>Dowód lematu:</u>

funkcja $\log t$ - wklęsła tzn. $\log\left(\sum_i w_i t_i\right) \geq \sum_i w_i \log t_i$

$w_i \geq 0 \quad \sum w_i = 1$

$$t_i = \frac{1}{p_1(x)} \qquad w_i = p_1(x)$$

$$\sum_x \underbrace{p_1(x)}_{w_i} \log \underbrace{\frac{p_2(x)}{p_1(x)}}_{t_i} \leq \log \sum_x p_2(x) = 0$$

czyli $$\sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)} \geq 0$$ $\blacksquare$

<u>Dowód twierdzenia</u>

Założenia techniczne: • istnieją 1-sza, 2-ga pochodne $\log p_\Theta(x)$

• $\left\langle \frac{\partial}{\partial \Theta} \log p_\Theta(x) \right\rangle = 0$

• asymptotyczna nieobciążoność

Niech $\tilde{\Theta}$ - będzie naszym estymatorem, funkcja wiarygodności dla niego:

$$\frac{1}{N}\log p_{\tilde\theta}^{(N)}(x^{(N)}) = \frac{1}{N}\sum_i \log p_{\tilde\theta}(x_i)$$

z prawa wielkich liczb, to zbiega do średniej czyli

$$\frac{1}{N}\log p_{\tilde\theta}^{(N)}(x^{(N)}) \xrightarrow{N\to\infty} \int dx\, p_{\theta_0}(x)\log p_{\tilde\theta}(x)$$

prawdziwa wartość

z lematu

$$\leq \int dx\, p_{\theta_0}(x)\log p_{\theta_0}(x)$$

Czyli maximum $\frac{1}{N}\log p_{\tilde\theta}^{(N)}(x^{(N)})$ dla $N\to\infty$

będzie odpowiadało $\tilde\theta_{ML} = \theta_0$

– asymptotyczna nieobciążoność $\tilde\theta_{ML}$

• asymptotyczna efektywność

Twierdzenie o wart. średniej | dla ustalenia uwagi: $\theta_0 < \tilde\theta$

$$\frac{\left.\frac{\partial \log p_\theta^{(N)}(x)}{\partial\theta}\right|_{\theta=\tilde\theta} - \left.\frac{\partial\log p_\theta^{(N)}(x)}{\partial\theta}\right|_{\theta=\theta_0}}{\tilde\theta-\theta_0} = \left.\frac{\partial^2\log p^{(N)}(x)}{\partial\theta^2}\right|_{\theta=\bar\theta}$$

gdzie $\theta_0 < \bar\theta < \hat\theta$

Jeśli $\tilde\theta = \tilde\theta_{ML}$ to $\left.\frac{\partial\log p_\theta^{(N)}(x)}{\partial\theta}\right|_{\theta=\tilde\theta_{ML}} = 0$

$$(*)\quad \left.\frac{\partial\log p_\theta^{(N)}(x)}{\partial\theta}\right|_{\theta=\theta_0} = \left.\frac{\partial^2\log p_\theta^{(N)}(x)}{\partial\theta^2}\right|_{\theta=\bar\theta}(\theta_0-\tilde\theta_{ML})$$

Weźmy:

$$\left.\frac{\partial^2\log p_\theta^{(N)}(x)}{\ldots}\right| = N\frac{1}{N}\sum_i \left.\frac{\partial^2\log p_\theta(x_i)}{\ldots}\right|$$

$$\left.\frac{\partial^2 \log p_\theta^{(N)}(x)}{\partial \theta^2}\right|_{\theta=\bar\theta} \qquad \left.\frac{\partial^2 \log p_\theta^{(N)}(x)}{\partial \theta^2}\right|_{\theta=\bar\theta}$$

Wiemy że dla $N \to \infty$ $\tilde{\theta}_{ML} \to \theta$ a więc i $\bar\theta$

$$= N \frac{1}{N} \sum \left.\frac{\partial^2 \log p_\theta(x_i)}{\partial \theta^2}\right|_{\theta=\theta_0} \xrightarrow{N} N \left\langle \left.\frac{\partial^2 \log p_\theta(x)}{\partial \theta^2}\right|_{\theta=\theta_0} \right\rangle$$

$$= -N F(\theta_0)$$

Z kolei

$$\left.\frac{\partial \log p_\theta^{(N)}(x)}{\partial \theta}\right|_{\theta=\theta_0} = \sqrt{N} \underbrace{\frac{1}{\sqrt{N}} \sum_i \left.\frac{\partial \log p_\theta(x_i)}{\partial \theta}\right|_{\theta=\theta_0}}_{\xi}$$

ξ - zmienna losowa będąca sumą niezależnych zmiennych losowych

$$\langle \xi \rangle = 0$$

$$\langle \xi^2 \rangle = \left\langle \left(\sum_i \frac{\partial \log p_\theta(x_i)}{\partial \theta} \right)^2 \right\rangle =$$

$$= \sum_i \left\langle \left(\frac{\partial \log p_\theta(x_i)}{\partial x_i}\right)^2 \bigg|_{\theta=\theta_0} \right\rangle = F(\theta_0)$$

$$\xi \overset{N\to\infty}{\sim} \mathcal{N}(0, F(\theta_0)^{-1})$$

Czyli:

$$\tilde{\theta}_{ML} - \theta_0 = \frac{\left.\frac{\partial \log p_\theta^{(N)}(x)}{\partial \theta}\right|_{\theta=\theta_0}}{\left.\frac{\partial^2 \log p_\theta^{(N)}(x)}{\partial \theta^2}\right|_{\theta=\bar\theta}} \overset{N\to\infty}{\sim} \mathcal{N}\left(0, \frac{N F(\theta_0)}{N^2 F^2(\theta_0)}\right)$$

$$\left(\tilde{\theta}_{ML} - \theta_0\right) \overset{N\to\infty}{\sim} \mathcal{N}\left(\theta, \frac{1}{NF}\right)$$

<u>Uwaga</u>: z góry nie wiadomo jak duże

N trzeba by więc ale zwykle coś < 1000 wystarczy.

Przykład ⌂ Max-likelihood dla estymacji φ

jeśli $p_0 = \sin^2\frac{\varphi}{2}$ $p_1 = \cos^2\frac{\varphi}{2}$ ale jakie

N odstęp od C-R mniejszy niż 1 %