3. Podejście Bayesowskie

dotychczas myśleliśmy tak: $p_\theta(x)$

$\uparrow$ pewien określony nieznany parametr

i mówiliśmy o rodzinie rozkładów prawdopodobieństwa

teraz myślimy tak $p(x|\theta)$

a θ jest zmienną losową z pewnym rozkładem a priori $p(\theta)$. Zalecane podejście jeśli chcemy poprawić estymację wykorzystując wiedzę a priori albo gdy nie potrafimy znaleźć opt. estymatorów (nie istnieją...) w podejściu klasycznym

3.1 Optymalny Estymator Bayesowski

W podejściu klasycznym minimalizowaliśmy:

$\sigma^2\tilde{\theta} = \int dx\, (\tilde{\theta}(x)-\theta)^2 p_\theta(x)$ i wymagaliśmy

nieobciążoności estymatora

W podejściu Bayesowskim minimalizujemy średnią wariancję:

$$\sigma^2\tilde{\theta} = \int dx\, d\theta\, (\tilde{\theta}(x)-\theta)^2 \underbrace{p(x|\theta)\, p(\theta)}_{p(x,\theta)}$$

Korzystając z wzoru Bayesa $p(x,\theta)=p(\theta|x)\,p(x)$:

$$= \int \left[\int d\theta \left[(\tilde{\theta}(x)-\theta)^2 p(\theta|x)\right]\right] p(x)\, dx$$

Ponieważ $p(x) \geq 0$ szukamy optymalnego estymatora oznacza znalezienie dla każdego x takiego $\tilde{\theta}(x)$ który minimalizuje $\int dx\,(\tilde{\theta}(x)-\theta)^2 p(\theta|x)$

$$\frac{d}{d\tilde{\theta}}\int d\theta\,(\tilde{\theta}-\theta)^2 p(\theta|x) = \int d\theta\, 2(\tilde{\theta}-\theta)\, p(\theta|x) = 0$$

$$\Downarrow$$

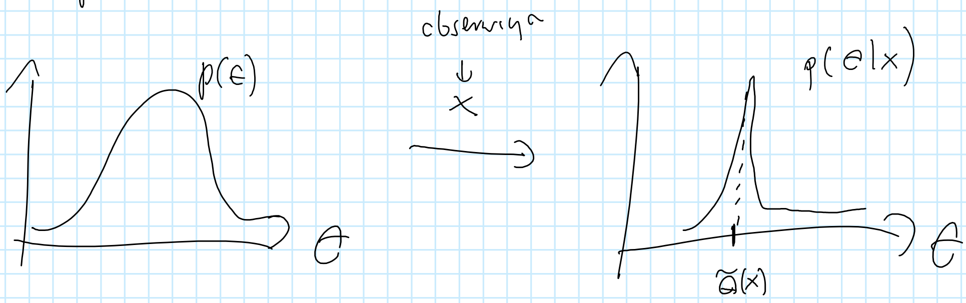
$$\tilde{\theta} = \int d\theta\, \theta\, p(\theta|x)$$

Optymalny estymator: Bayesowski:

$$\boxed{\tilde{\theta}(x) = \langle\theta\rangle_{p(\theta|x)}}$$

wartość oczekiwana rozkładu a posteriori

$$\underbrace{p(\theta|x)}_{\text{a posteriori}} = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)\,\overbrace{p(\theta)}^{\text{a priori}}}{p(x)}$$



Dla optymalnego estymatora:

$$\sigma^2\tilde{\theta} = \int dx\, d\theta\,(\theta-\tilde{\theta}(x))^2 p(\theta|x)\,p(x)\,dx =$$

$$= \int dx \left[\int d\theta\,(\theta-\langle\theta\rangle_{p(\theta|x)})^2 p(\theta|x)\right] p(x)\,dx$$

$$= \int dx\, \sigma^2\theta|_{p(\theta|x)}\, p(x)\,dx$$

<u>Intuicja</u>:

W ogólności $\sigma^2\tilde{\theta}$ będzie zależeć od wiedzy a priori ale im więcej danych tym wpływ wiedzy a priori słabszy;

$$p(\theta|x) \propto \underset{\substack{\uparrow \\ \text{szybko zmienne} \\ \text{z } \theta}}{p(x|\theta)} \; \underset{\substack{\uparrow \\ \text{wolno zmienne} \\ \text{z } \theta}}{p(\theta)} \simeq p(x|\theta)$$

Jeśli danych mało estymator będzie dociągany w kierunku średniej rozkładu apriori;

największy problem podejścia Bayesowskiego: to wybór rozkładu apriori!

Uwaga:

ponieważ mamy dane $p(x|\theta)$, $p(\theta)$, aby obliczyć $p(\theta|x)$ musimy obliczyć $p(x) = \int d\theta \, p(x|\theta) p(\theta)$

$$p(\theta|x) = \frac{p(x|\theta) \; p(\theta)}{\int d\theta' \, p(x|\theta') p(\theta')}$$

Nie zawsze łatwo wyliczyć całkę, ale wyrażenie w mianowniku to po prostu normalizacja rozkładu $p(\theta|x)$, bo nie zależy od θ.

Przykład

$$x_i \sim N(\theta, \sigma^2) \qquad i = 1 \dots N$$

$$p(x|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_i (x_i - \theta)^2}$$

Weźmy gaussowski rozkład a priori

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{1}{2\sigma_\theta^2}(\theta - \underset{\substack{\uparrow \\ \text{średnia a priori}}}{\mu_\theta})^2}$$

$$p(x|\theta)\cdot p(\theta) =$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \cdot e^{-\frac{1}{2}\overbrace{\left[\frac{1}{\sigma^2}\sum_i (x_i-\theta)^2 + \frac{1}{\sigma_\theta^2}(\theta-\mu_\theta)^2\right]}^{G(\theta)}}$$

Rozkład a posteriori:

$$p(\theta|x) = \frac{e^{-\frac{1}{2}G(\theta)}}{\int d\theta' \, e^{-\frac{1}{2}G(\theta')}\, d\theta'}$$

Przepiszmy $G(\theta)$:

$$G(\theta) = \theta^2\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right) - 2\theta\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_\theta}{\sigma_\theta^2}\right)$$

$$+ \underbrace{\frac{1}{\sigma^2}\sum_i x_i^2 + \frac{1}{\sigma_\theta^2}\mu_\theta^2}_{\text{te nie zależą od } \theta}$$

więc się skrócą

Możemy dodawać i odejmować człony niezależne od θ

Zdefiniujmy
$$\sigma_{\theta|x}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

$$\mu_{\theta|x} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\sigma_{\theta|x}^2$$

Wtedy z dokładnością do członów nie zależnych od θ

$$G(\theta) = \frac{1}{\sigma_{\theta|x}^2}\left(\theta - \mu_{\theta|x}\right)^2 + \ldots$$

Czyli:

$$1 \quad e^{-\frac{1}{2\sigma_{\theta|x}^2}\left(\theta-\mu_{\theta|x}\right)^2}$$

$$p(\theta|x) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}} e^{-\frac{1}{2\sigma_{\theta|x}^2}(\theta - \mu_{\theta|x})^2}$$

Rozkład a posteriori też gaussowski (przeliczenia wcześniej nie były problemem)

Widzimy że jego wariancja

$$\sigma_{\theta|x}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

będzie tym większa im więcej prób a wpływ rozkładu a priori będzie malał im większe N. Średnia z kolei

$$\mu_{\theta|x} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\sigma_{\theta|x}^2 =$$

$$= \underbrace{\frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_\theta^2}}}_{\alpha}\bar{x} + \underbrace{\frac{\frac{1}{\sigma_\theta^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_\theta^2}}}_{1-\alpha}\mu_\theta =$$

$$= \alpha \cdot \bar{x} + (1-\alpha)\mu_\theta$$

wkład od danych

obciążenie ze średniej a priori

Dla dużych N rozkład a priori nie ma znaczenia: $\alpha \to 1$, im α bliższe 0 tym większy wpływ rozkładu a priori.

Dla tego problemu optymalny estymator jest trywialny do znalezienia (średnia a posteriori)

$$\tilde{\theta}(x) = \mu_{\theta|x}$$

A średnia wariancja estymatora

$$\Delta^2\tilde{\theta} = \int dx\, \Delta^2\theta|_{p(\theta|x)}\, p(x) = \int dx\, \sigma^2_{\theta|x}\, p(x) = \frac{1}{\frac{N}{\sigma^2}+\frac{1}{\sigma_\theta^2}}$$

W granicy $N \to \infty$ $\quad \tilde{\theta}(x) = \bar{x} \quad \Delta^2\tilde{\theta} = \frac{\sigma^2}{N}$

Przyjemnie jest pracować z rozkładami a priori dla których a posteriori łatwe do liczenia całkowania itp. Tak jest dla gaussowskich rozkładów a priori i gaussowskich procesów

3.2

3.3 Bayesian Cramer-Rao bound + bernstein von Mises etc….

Jakie są związki między podejściem Bayesowskim a podejściem przez informację Fishera.

Wyprowadźmy analog nierówności C-R dla podejścia Bayesowskiego:

$$\Delta^2\tilde{\theta} = \int dx\, d\theta\; p(x|\theta)\, p(\theta)\, \left(\tilde{\theta}(x) - \theta\right)^2 = \int dx\, d\theta\; f^2(x,\theta)$$

$$f(\theta,x) = \sqrt{p(x|\theta)\, p(\theta)}\; \left(\tilde{\theta}(x) - \theta\right)$$

Zdefiniujmy:

$$g(\theta,x) = \sqrt{\frac{p(\theta)}{p(x|\theta)}}\; \frac{dp(x|\theta)}{d\theta} + \sqrt{\frac{p(x|\theta)}{p(\theta)}}\; \frac{dp(\theta)}{d\theta} = \frac{1}{\sqrt{p(x|\theta)\, p(\theta)}}\; \frac{d\, p(x|\theta)\, p(\theta)}{d\theta}$$

$$g(\theta,x)=\sqrt{\frac{p(\theta)}{p(x|\theta)}}\frac{dp(x|\theta)}{d\theta}+\sqrt{\frac{p(x|\theta)}{p(\theta)}}\frac{dp(\theta)}{d\theta}=\frac{1}{\sqrt{p(x|\theta)p(\theta)}}\frac{d\,p(x|\theta)p(\theta)}{d\theta}$$

$$\int g(\theta,x)^2\,d\theta\,dx=\int d\theta\,dx\;\frac{p(\theta)}{p(x|\theta)}\left(\frac{dp(x|\theta)}{d\theta}\right)^2+\frac{p(x|\theta)}{p(\theta)}\left(\frac{dp(\theta)}{d\theta}\right)^2$$

$$\underbrace{+\int 2\,\frac{dp(x|\theta)}{d\theta}\,\frac{dp(\theta)}{d\theta}}_{0}=\underbrace{\int p(\theta)\,d\theta\,F_\theta}_{F}+\underbrace{\int d\theta\,\frac{1}{p(\theta)}\left(\frac{dp(\theta)}{d\theta}\right)^2}_{I}$$

$$\int p(\theta,x)\,g(\theta,x)=\int p(\theta)\left(\tilde{\theta}(x)-\theta\right)\frac{dp(x|\theta)}{d\theta}+p(x|\theta)\left(\tilde{\theta}(x)-\theta\right)\frac{dp(\theta)}{d\theta}$$

$$=\int\left(\tilde{\theta}(x)-\theta\right)\frac{d\left(p(x|\theta)p(\theta)\right)}{d\theta}=\int dx\,\tilde{\theta}(x)\,\underbrace{p(x,\theta)\Big|_{\theta_-}^{\theta_+}}_{=0}-\int\theta\,\frac{dp(\theta)}{d\theta}\,d\theta$$

$$=\underbrace{-\theta\,p(\theta)\Big|_{\theta_-}^{\theta_+}}_{=0}+\int p(\theta)\,d\theta=1$$

zakładamy że na brzegach znika $p(\theta)$

$$\Downarrow$$

$$\boxed{\sigma^2_{\tilde{\theta}_B}\geq\frac{1}{F+I}}$$

wkład Fisher

wkład od wiedzy a priori

Przykład

$x_i\sim N(\theta,\sigma^2)$ $\quad i=1\dots N$

$$p(x|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\theta)^2}$$

Weźmy gaussowski rozkład a priori

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{1}{2\sigma_\theta^2}(\theta-\mu_\theta)^2}$$

↑ średnia a priori

$$F = N \cdot \frac{1}{\sigma^2}$$

$$I = \int \sqrt{2\pi\sigma_\theta^2}\, e^{+\frac{1}{2\sigma^2}(\theta-\mu_B)^2} \frac{(\theta-\mu_B)^2}{\sigma_\theta^4} \left[\frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{1}{2\sigma_\theta^2}(\theta-\mu_B)^2}\right]^2$$

$$= \frac{1}{\sigma_\theta^2}$$

$$\Delta^2\tilde{\theta} \geq \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

Widzimy że dla nieobciążonego przypadku mamy wysycenie bo wcześniej uzyskaliśmy równość. Ogólnie tylko wtedy mamy wysycenie Θ..

Bernstein-von-Mises theorem

(tak naprawdę w pewnym szczególnym przypadku)

Przy pewnych założeniach o regularności $p(\theta)$, $p(x|\theta)$

Dla N, powtórzeń eksperymentu, $(x_1 \dots x_N)$, estymatory średniej wartości $\tilde{\theta}(\tilde{x}) = \int d\theta \, p(\theta | x_1 \dots x_N)\, \theta$

i $\tilde{\theta}$ dąży dla Gaussa skupionego wokół prawdziwej wartości:

$$\sqrt{N}\left(\tilde{\theta}_B - \theta_0\right) \xrightarrow{N \to \infty} N\left(0, \frac{1}{F_{\theta_0}}\right)$$

Nie ma znaczenia prior ...