Bayesian Model Selection

① D $\quad M_1, M_2, \dots M_K$ $\quad \mathcal{M} \overset{def}{=} p(\bar\theta), p(D|\bar\theta)$

$p(\bar\theta|\mathcal{M}),\ p(D|\bar\theta,\mathcal{M})$

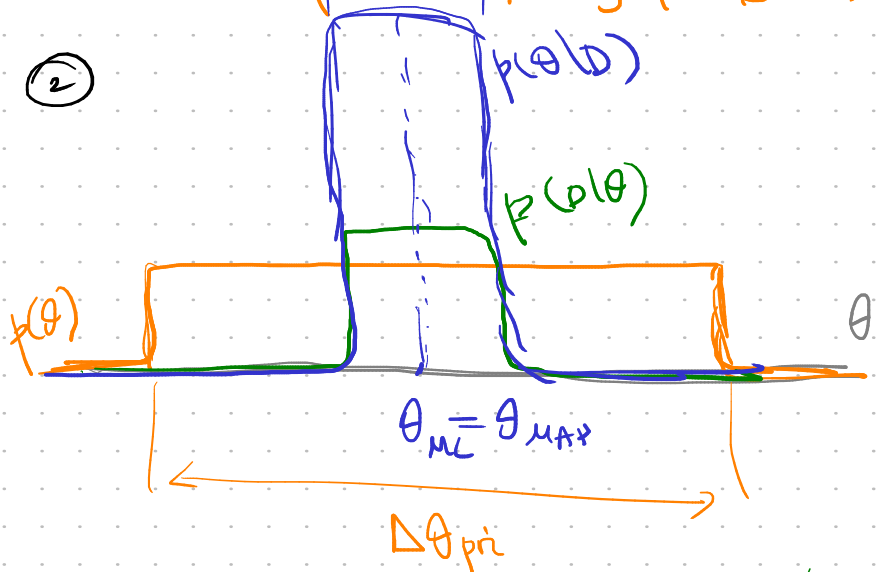
$$p(M_k|D) = \frac{p(M_k)\,p(D|M_k)}{p(D)}$$

$p(M_k) = 1/K$

$\underset{k}{\text{argmax}}\ p(D|M_k)$

$$p(\bar\theta_k|D,M_k) = \frac{p(\bar\theta_k|M_k)\cdot p(D|\bar\theta_k,M_k)}{p(D|M_k)}$$

$$p(D|M_k) = \int p(\bar\theta_k|M_k)\,p(D|\bar\theta_k,M_k)\,d\bar\theta_k$$

② 

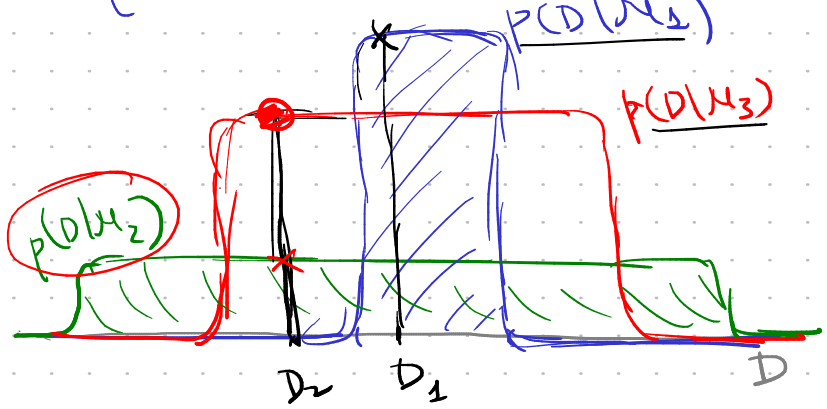
$$p(D) = \int p(\bar\theta)\,p(D|\bar\theta)\,d\bar\theta = \int_{(\Delta\theta_{pri})^d} \left(\frac{1}{\Delta\theta_{pri}}\right)^d p(D|\bar\theta)\,d\bar\theta =$$

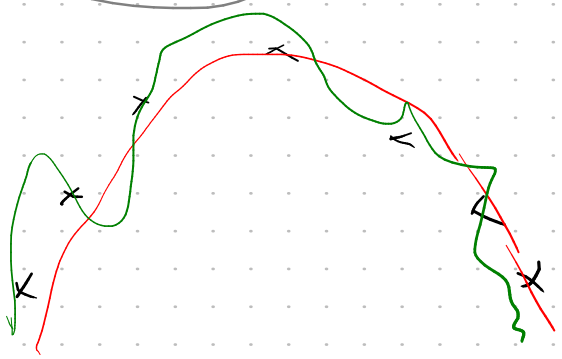
$$= \int_{(\Delta\theta_{post})^d} \frac{1}{(\Delta\theta_{pri})^d}\,p(D|\bar\theta_{MAP})\,d\bar\theta = p(D|\bar\theta_{MAP})\cdot\left(\frac{\Delta\theta_{post}}{\Delta\theta_{pri}}\right)^d$$

information criterion

$$\log p(D) = \log p(D|\theta_{MAP}) - d\log\frac{\Delta\theta_{pri}}{\Delta\theta_{post}}$$

$\theta_{MAP} = \theta_{ML}$

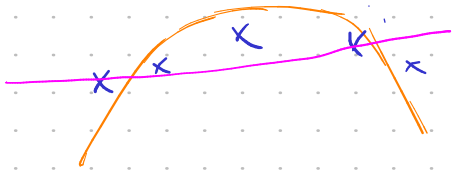




③ Sanity check

$D \sim M_{true}$ $\quad p(D|M_{true}) \geq p(D|M)$

$$\mathbb{E}_{D\sim p(D|M_{true})}\left[\ln \frac{p(D|M_{true})}{p(D|M)}\right] =$$

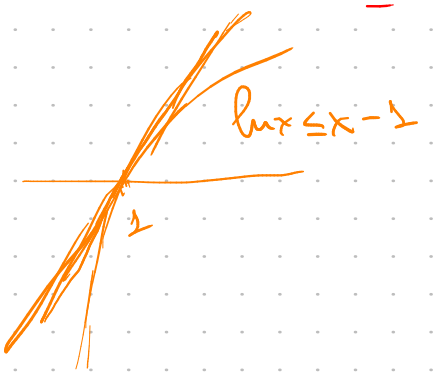


$$= \int p(D|M_{true}) \ln \frac{p(D|M_{true})}{p(D|M)}\, dD = KL(p(D|M_{true}) \| p(D|M)) \geq 0$$

$$KL(p\|q) = \int p(\bar{x}) \ln \frac{p(\bar{x})}{q(\bar{x})} d\bar{x} = \text{Kullback-Leibler divergence}$$

$$= \int p(\bar{x}) \ln p(\bar{x})\, dx - \int p(\bar{x}) \ln q(\bar{x}) d\bar{x}$$

$$= H(p) - H(p,q) \qquad KL(q\|p) = \int q \ln \frac{q}{p} dx$$



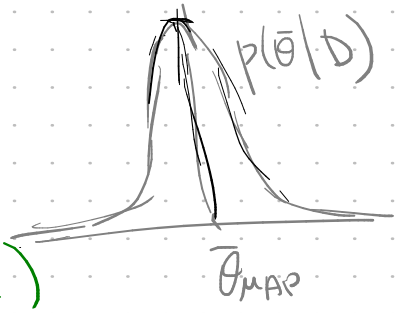
$$= -\int p \ln \frac{q}{p} d\bar{x} \geq -\int p \left(\frac{q}{p} - 1\right) d\bar{x}$$

$$= -\left(\int q\, d\bar{x} - \int p\, d\bar{x}\right) = 0$$

④ BIC – Bayesian information criterion

$$p(D) = \int p(\bar{\theta}) p(D|\bar{\theta}) d\bar{\theta}$$

$\approx N(\bar{\theta}|\bar{\theta}_{MAP}, \bar{\mu}, \ldots)$



$$p(D) = \int \underbrace{f(\bar{\theta})}_{\text{const}} d\theta \approx \int f(\bar{\theta}_{MAP}) \cdot e^{-\frac{1}{2}(\bar{\theta}-\bar{\theta}_{MAP})^T A (\bar{\theta}-\bar{\theta}_{MAP})} d\bar{\theta}$$

$$= f(\bar{\theta}_{MAP}) \cdot \sqrt{\frac{(2\pi)^d}{\det A}}, \quad \text{where } A = -\nabla\nabla \ln f(\bar{\theta})\Big|_{\bar{\theta}_{MAP}}$$

$$\ln f(\bar{\theta}) \approx \ln f(\bar{\theta}_{MAP}) + (\bar{\theta}-\bar{\theta}_{MAP})^T \nabla \ln f + \frac{1}{2}(\bar{\theta}-\bar{\theta}_{MAP})\left(-\frac{\partial^2 \ln f}{\partial\theta_i \partial\theta_j}\right)(\bar{\theta}-\bar{\theta}_{MAP})$$

$$\ln p(D) \approx \ln p(D|\bar{\theta}_{MAP}) + \ln p(\bar{\theta}_{MAP}) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \det A$$

Occam's factor

$N = |D|$

AIC - Akaike IC

BIC

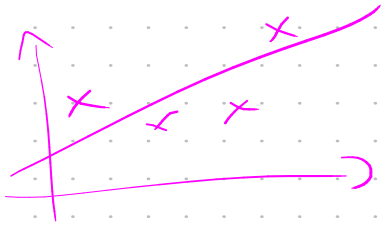
$$\ln p(D) \approx \ln p(D|\bar{\theta}_{MAP}) - \frac{1}{2}\cdot d \cdot \ln N$$

(5) Empirical Bayes

hyperparameters

$$p(y|\bar{x},\bar{w}) = \mathcal{N}(y|\bar{w}^T\bar{x}, \sigma^2)$$

$$p(\bar{w}) = \mathcal{N}(\bar{w}|\bar{0}, \sigma_0^2 \cdot \bar{I})$$



$$\log p(\bar{w}|\bar{y},X) = \text{const} - \frac{1}{2\sigma^2}\sum_n (y_n - \bar{x}_n^T\bar{w})^2 - \frac{1}{2\sigma_0^2}\alpha\,\bar{w}^T\bar{w}$$

$$\log p(\bar{w}|\bar{y},X) = \text{const} - \frac{\beta}{2}\sum_n (-)^2 - \frac{\alpha}{2}\bar{w}^T\bar{w}$$

MLE - II

$$p(\bar{w}|D,\alpha,\beta) = \frac{p(\bar{w}|\alpha)\,p(D|\bar{w},\beta)}{p(D|\alpha,\beta)} \xrightarrow[\alpha,\beta]{} \max$$

evidence function
marginal likelihood

$$p(\bar{y}|X,\alpha,\beta) = \int p(\bar{w}|\alpha)\,p(\bar{y}|X,\bar{w},\beta)\,d\bar{w}$$

$$= \int \left(\frac{\alpha}{2\pi}\right)^{d/2} e^{-\frac{\alpha}{2}\bar{w}^T\bar{w}} \cdot \left(\frac{\beta}{2\pi}\right)^{N/2} e^{-\frac{\beta}{2}\sum_n (y_n - \bar{w}^T\bar{x}_n)^2} d\bar{w}$$

$E(\bar{w})$

$$-\frac{\alpha}{2}\bar{w}^T\bar{w} - \frac{\beta}{2}(\bar{y}-X\bar{w})^T(\bar{y}-X\bar{w}) =$$

$$= -\frac{\alpha}{2}\bar{w}^T\bar{w} - \frac{\beta}{2}(\bar{y}^T\bar{y} - 2\bar{w}^TX^T\bar{y} + \bar{w}^TX^TX\bar{w}) =$$

$$= -\frac{1}{2}\bar{w}^T\underbrace{(\beta X^TX + \alpha\cdot I)}_{A}\bar{w} + \beta\,\bar{w}^T\underbrace{X^T\bar{y}}_{A\bar{\mu}} - \frac{\beta}{2}\bar{y}^T\bar{y} =$$

$$= -\frac{1}{2}(\bar{w}-\bar{\mu})^T A(\bar{w}-\bar{\mu}) - \frac{\beta}{2}\bar{y}^T\bar{y} + \frac{1}{2}\bar{\mu}^T A\bar{\mu}, \text{ где}$$

$$A = \beta \cdot X^T X + \alpha I = \Sigma_N^{-1}, \quad \bar{\mu} = \bar{\mu}_N = \beta \cdot A^{-1} X^T \bar{y}$$

«$A\bar{\mu}/\beta$»

$$E(\bar{\mu}) = -\frac{\alpha}{2}\bar{\mu}^T\bar{\mu} - \frac{\beta}{2}(\bar{y}^T\bar{y} - 2\bar{\mu}^T X^T\bar{y} + \bar{\mu}^T X^T X\bar{\mu}) =$$

$$-\frac{\beta}{2}\bar{y}^T\bar{y} + \frac{\beta}{2}\bar{\mu}^T X^T X\bar{\mu} + \frac{\alpha}{2}\bar{\mu}^T\bar{\mu} =$$

$$= -\frac{\beta}{2}(\bar{\mu}^T X^T X\bar{\mu} + \bar{y}^T\bar{y} - 2\bar{\mu}^T X^T\bar{y}) - \beta\bar{\mu}^T X^T\bar{y} + \frac{\alpha}{2}\bar{\mu}^T\bar{\mu} + \beta\bar{\mu}^T X^T X\bar{\mu}$$

$$= \quad - \bar{\mu}^T A\bar{\mu} + \frac{\alpha}{2}\bar{\mu}^T\bar{\mu} + \beta\bar{\mu}^T X^T X\bar{\mu}$$

$$= \quad - \bar{\mu}^T(\beta X^T X + \alpha I)\bar{\mu} + \frac{\alpha}{2}\bar{\mu}^T\bar{\mu} + \beta\bar{\mu}^T X^T X\bar{\mu}$$

$$= \quad - \frac{\alpha}{2}\bar{\mu}^T\bar{\mu}$$

$$E(\bar{w}) = E(\bar{\mu}) - \frac{1}{2}(\bar{w}-\bar{\mu})^T A(\bar{w}-\bar{\mu}) \qquad \alpha, \beta \to \max$$

$$\log p(D|\alpha,\beta) = E(\bar{\mu}) + \frac{d}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log 2\pi - \frac{1}{2}\log\det A$$

$$-\frac{d}{2}\log 2\pi \qquad +\frac{d}{2}\log 2\pi$$

$\log\det A$

$A = \alpha I + \beta(X^T X)$

λ_i — собств. числа $\beta(X^T X)$

$\lambda_i + \alpha$ — с.ч. A

$\log\det A = \sum_{i=1}^{d}\log(\lambda_i + \alpha)$

$+\delta\beta \quad +\alpha\Sigma \quad \alpha\bar{v}_i$

$\beta(X^T X)\cdot\bar{v}_i = \lambda_i\bar{v}_i$

$\lambda_i \alpha \beta$

$\frac{\partial\lambda_i}{\partial\beta} = D_i = \frac{\lambda_i}{\beta} \qquad \lambda_i = D_i\beta$

$$\frac{\partial\log p(D|\alpha,\beta)}{\partial\alpha} = \frac{d}{2\alpha} - \frac{1}{2}\bar{\mu}^T\bar{\mu} - \frac{1}{2}\sum_i \frac{1}{\lambda_i+\alpha} = 0$$

$$\alpha - \bar{\mu}^T\bar{\mu} + \alpha \sum \frac{1}{\lambda_i + \alpha} = d \qquad \bar{\mu} = \bar{\mu}(\alpha, \beta)$$

$$\alpha^{(k+1)} = \frac{1}{\bar{\mu}^T\bar{\mu}} \cdot \sum_{i=1}^{d}\left(1 - \frac{\alpha}{\lambda_i + \alpha}\right) = \frac{1}{\bar{\mu}^{(k)T}\bar{\mu}^{(k)}} \cdot \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \alpha^{(k)}}$$

$$\frac{\partial \log p(D|\alpha,\beta)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2}\cdot\sum_n (y_n - \bar{x}_n^T\bar{\mu})^2 - \frac{1}{2}\sum_i \frac{\lambda_i/\beta}{\lambda_i + \alpha} = 0$$

$$\beta^{(k+1)} = \frac{N - \sum_i \frac{\lambda_i}{\lambda_i + \alpha^{(k)}}}{\sum_n (y_n - \bar{\mu}^{(k)T}\bar{x}_n)^2}$$