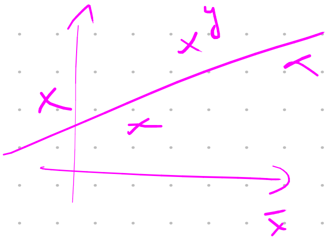
Empirical Bayes



$1/\beta$

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

$$p(\bar w) = \mathcal{N}(\bar w|\bar 0, \alpha I)$$

$$\log p(\bar w|X,\bar y) = const + \log p(\bar w) + \log p(\bar y|\bar w, X) =$$

$$= const - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \bar w^T\bar x_n)^2 - \frac{\alpha}{2}\bar w^T\bar w \xrightarrow[\bar w]{} max$$

$$= const - \frac{\beta}{2}\sum_n (y_n - \bar w^T\bar x_n)^2 - \frac{\alpha}{2}\bar w^T\bar w$$

hyperparameters ?

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

$$p(\bar\theta|D,\bar\alpha) = \frac{p(\bar\theta|\bar\alpha)\cdot p(D|\bar\theta,\bar\alpha)}{p(D|\bar\alpha)}$$

evidence function

marginal likelihood

$$p(D|\bar\alpha) = \int p(D,\bar\theta|\bar\alpha)\,d\bar\theta =$$

$$= \int p(\bar\theta|\bar\alpha)\,p(D|\bar\theta,\bar\alpha)\,d\bar\theta$$

$$\approx \frac{1}{N}\sum_r p(D|\bar\theta_r,\bar\alpha)$$

$\theta_r \sim p(\bar\theta|\bar\alpha)$

Empirical Bayes

Type II Max likelihood estimation — MLE-II

$$p(D|\bar\alpha) \xrightarrow[\bar\alpha]{} max$$

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

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

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

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

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

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

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

$$\int e^{-\frac{1}{2}(\bar{w}-\bar{\mu}_N)^T A (\bar{w}-\bar{\mu}_N)} d\bar{w} = (2\pi)^{d/2} \cdot \sqrt{\det A^{-1}} = \sqrt{\frac{(2\pi)^d}{\det A}}$$

$$p(\bar{x} \mid \bar{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2} \cdot \sqrt{\det \Sigma}} \cdot e^{-\frac{1}{2}(\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu})}$$

$$p(\bar{y} \mid X, \alpha, \beta) = \left(\frac{\alpha}{2\pi}\right)^{d/2} \left(\frac{\beta}{2\pi}\right)^{N/2} \cdot \sqrt{\frac{(2\pi)^d}{\det A}} \, e^{\frac{1}{2}\bar{\mu}_N^T A \bar{\mu}_N - \frac{\beta}{2}\bar{y}^T\bar{y}} \xrightarrow[\alpha,\beta]{} \max$$

$$\log p(\bar{y} \mid X, \alpha, \beta) = \text{Const} + \frac{d}{2}\log\alpha + \frac{N}{2}\log\beta -$$

$$- \frac{1}{2}\log\det A \underbrace{- \frac{\beta}{2}\bar{y}^T\bar{y} + \frac{1}{2}\bar{\mu}_N^T A \bar{\mu}_N} \xrightarrow[\alpha,\beta]{} \max$$

$$\beta\, \bar{\mu}_N^T X^T \bar{y} = \bar{\mu}_N^T \cdot A (\overbrace{A^{-1} \cdot \beta X^T \bar{y}}^{\bar{\mu}_N}) = \bar{\mu}_N^T A \bar{\mu}_N = \alpha \cdot \underline{\bar{\mu}_N^T \bar{\mu}_N + \beta \bar{\mu}_N X^T X \bar{\mu}_N}$$

$$-\frac{\beta}{2}\bar{y}^T\bar{y} + \frac{1}{2}\bar{\mu}_N^T A \bar{\mu}_N = \underline{-\frac{\beta}{2}\bar{y}^T\bar{y}} + \frac{\alpha}{2}\bar{\mu}_N^T \bar{\mu}_N + \underline{\frac{\beta}{2}\bar{\mu}_N X^T X \bar{\mu}_N}$$

$$-\alpha \cdot \bar{\mu}_N^T \bar{\mu}_N - \cancel{\beta \bar{\mu}_N X^T X \bar{\mu}_N}$$

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

$$+ \cancel{\beta \cdot \bar{\mu}_N X^T X \bar{\mu}_N} + \frac{\alpha}{2}\bar{\mu}_N^T \bar{\mu}_N =$$

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

$$E(\bar{w}) = E(\bar{\mu}_N) - \frac{1}{2}(\bar{w} - \bar{\mu}_N) A (\bar{w} - \bar{\mu}_N)$$

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

$$\log p(\bar{y} | X, \alpha, \beta) = E(\bar{\mu}_N) + \frac{d}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{1}{2}\underline{\log\det A} + \text{const}$$

D_i – c.ч. $(X^T X)$ $\quad \lambda_i = D_i \beta$

λ_i : c.ч. $\underline{\beta X^T X}$: $\quad (\beta X^T X) \cdot \bar{u}_i = \lambda_i \bar{u}_i$

c.ч. A : $\quad (\alpha I + \beta X^T X)\bar{u}_i = \lambda_i \bar{u}_i + \alpha \bar{u}_i = (\alpha + \lambda_i)\bar{u}_i$

$$\log\det A = \sum_i \log(\alpha + \lambda_i)$$

$$\frac{\partial \log\det A}{\partial \alpha} = \sum_i \frac{1}{\alpha + \lambda_i}$$

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

$$\frac{d}{\alpha} = \bar{\mu}_N^T \bar{\mu}_N + \sum \frac{1}{\alpha + \lambda_i}$$

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

$\alpha^{(0)}$

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

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

$$\frac{\partial \log\det A}{\partial \beta} = \sum_i \frac{\partial \log(\lambda_i + \alpha)}{\partial \beta} = \sum_i \frac{D_i}{\lambda_i + \alpha}$$

$$\frac{\partial \log p(\bar{y}\mid X,\alpha,\beta)}{\partial \beta} = -\frac{1}{2}(\bar{y}-X\bar{\mu}_N)^T(\bar{y}-X\bar{\mu}_N) + \frac{N}{2\beta} - \frac{1}{2}\sum_i \frac{D_i}{\lambda_i+\alpha} = 0$$

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

$$-\frac{1}{2\beta}\sum \frac{\lambda_i}{\lambda_i+\alpha}$$

$$\frac{N-\Sigma}{\beta} = (\ldots)^T(\ldots)$$

$$\beta^{(k+1)} = \frac{N - \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i+\alpha}}{(\bar{y}-X\bar{\mu}_N(\beta^{(k)}))^T(\bar{y}-X\bar{\mu}_N(\beta^{(k)}))}$$

$$\lambda_i = D_i \cdot \beta^{(k)}$$

Model selection

$D \qquad \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \ldots, \mathcal{M}_K$

$$\mathcal{M}_k : \bar{\theta}_k \qquad p(\bar{\theta}_k \mid D, \mathcal{M}_k) = \frac{p(\bar{\theta}_k\mid \mathcal{M}_k)\, p(D\mid\bar{\theta}_k,\mathcal{M}_k)}{p(D\mid\mathcal{M}_k)}$$

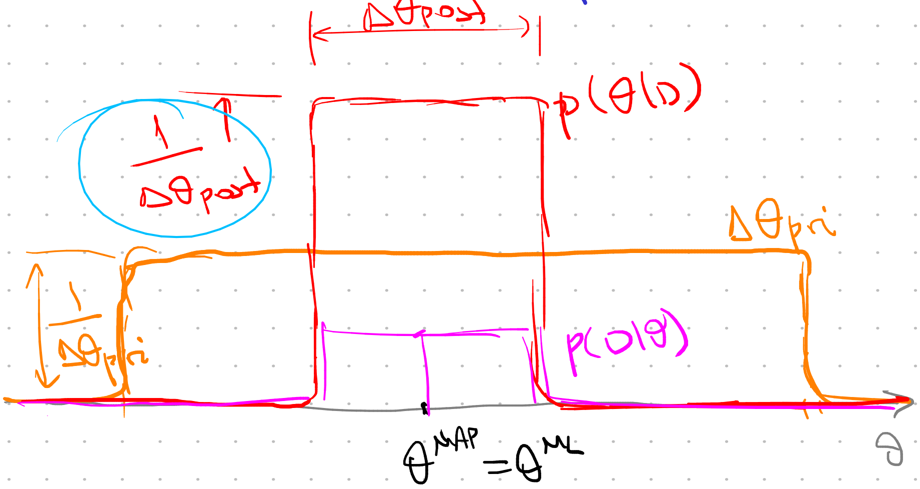
model evidence

unif

$$p(\mathcal{M}_k\mid D) \propto p(\mathcal{M}_k)\, p(D\mid\mathcal{M}_k) \propto p(D\mid\mathcal{M}_k)$$

$$\mathcal{M}_k : \bar{\theta}_k^{ML} = \arg\max p(D\mid\bar{\theta}_k,\mathcal{M}_k) \qquad p(D\mid\bar{\theta}_k^{ML},\mathcal{M}_k)$$

$$\bar{\theta}_k^{MAP} = \arg\max p(\bar{\theta}_k\mid\mathcal{M}_k)\, p(D\mid\bar{\theta}_k,\mathcal{M}_k) \qquad p(D\mid\bar{\theta}_k^{MAP},\mathcal{M}_k)$$



$$p(D\mid\theta) = \begin{cases} \frac{p(D\mid\theta^{MAP})}{\Delta\theta_{post}}, & \underline{\hspace{2em}} \\ 0, & \underline{\hspace{4em}} \end{cases}$$

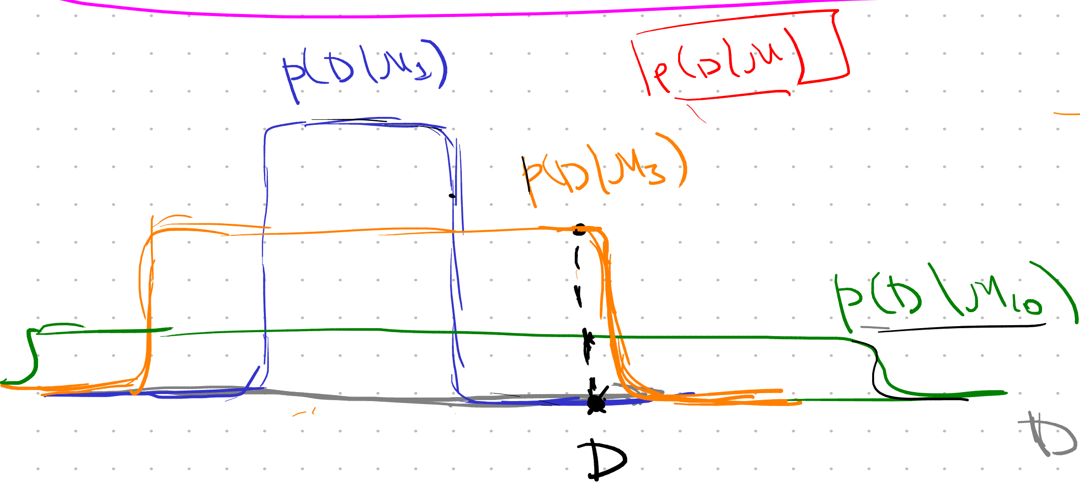
$$p(D) = \int p(\theta) \cdot p(D|\theta)\, d\theta =$$

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

$$p(D) = p(D|\overset{\theta^{ML}}{\overset{\shortparallel}{\theta^{MAP}}}) \cdot \frac{\Delta\theta_{post}}{\Delta\theta_{pri}}$$

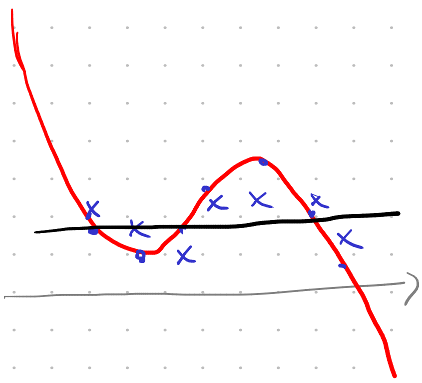
model selection criterion

$$\boxed{\ln p(D) = \ln p(D|\bar{\theta}^{MAP}) + M \cdot \ln \frac{\Delta\theta_{post}}{\Delta\theta_{pri}}}$$



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

$\cancel{\forall M \quad p(D|M_{true}) \geq p(D|M)\,?}$



$\underline{\mathbb{E}_{D\sim M_{true}}}$

$$\mathbb{E}_{D\sim p(D|M_{true})}\left[\log \frac{p(D|M_{true})}{p(D|M)}\right] = \int p(D|M_{true}) \cdot \log \frac{p(D|M_{true})}{p(D|M)}\, dD$$

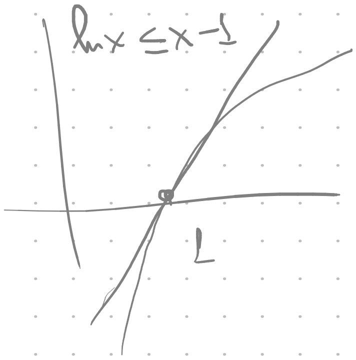
$$= KL\big(p(D|M_{true}) \,\|\, p(D|M)\big)$$

$$KL(p(\bar{x})\,\|\,q(\bar{x})) = \int p(\bar{x}) \ln \frac{p(\bar{x})}{q(\bar{x})}\, d\bar{x} =$$

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

$$= H(p,q) - H(p)$$

$KL(p\|q) \geq 0$?

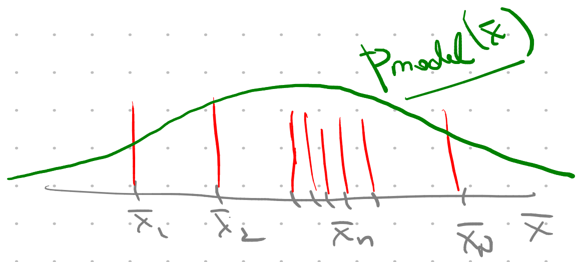


$$\int p(\bar{x}) \ln \frac{p(\bar{x})}{q(\bar{x})}\, d\bar{x} =$$

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

$$\geq -\int p(\bar{x}) \left(\frac{q(\bar{x})}{p(\bar{x})} - 1 \right) d\bar{x} = -\int q(\bar{x})\, dx + \int p(\bar{x})\, d\bar{x} = 0$$

D $\quad p_{data}(\bar{x}) \approx p_{model}(\bar{x}\,|\,\bar{\theta})$



$D = \{\bar{x}_1, \dots, \bar{x}_N\}$

$$p_{data}(\bar{x}) = Unif(D) = \frac{1}{N} \cdot \sum_{n=1}^{N} \delta(\bar{x} - \bar{x}_n)$$

$$KL(p_{data}\,\|\,p_{model}) = \int p_{data}(\bar{x}) \cdot \ln \frac{p_{data}(\bar{x})}{p_{model}(\bar{x})}\, d\bar{x} =$$

$$= -H(p_{data}) - \int p_{data}(\bar{x}) \ln p_{model}(\bar{x})\, d\bar{x} =$$

$$= -H(p_{data}) - \frac{1}{N} \underbrace{\sum_{n=1}^{N} \ln p_{model}(\bar{x}_n)}_{\parallel} \xrightarrow{\bar{\theta}} \min$$

$$\log p_{model}(D\,|\,\bar{\theta}) \xrightarrow{\bar{\theta}} \max$$

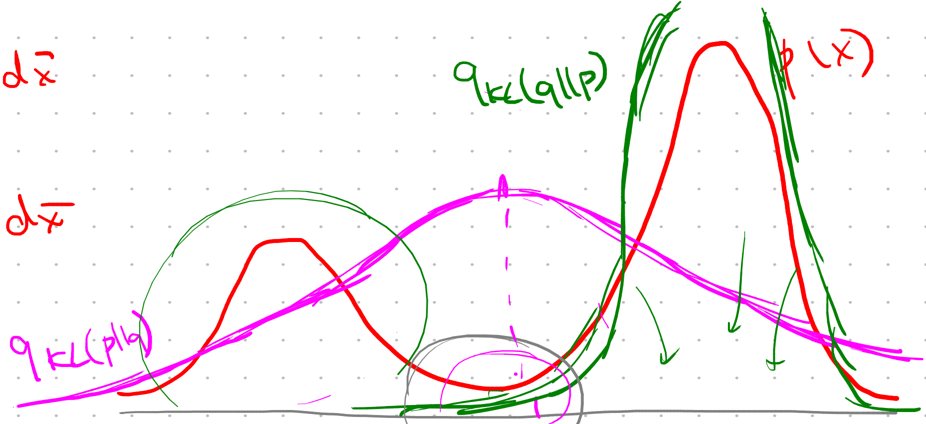
$p(\bar{\theta}|D) \propto p(\bar{\theta})\, p(D|\bar{\theta})$

$\to \bar{\theta}_{MAP}$

$\to p(\bar{x}|D) \approx \frac{1}{N} \sum p(\bar{x}|\bar{\theta})$

$$KL(p\|q) = \int p \ln \frac{p}{q}\, d\bar{x}$$

$$\rightarrow\ KL(q\|p) = \int q \ln \frac{q}{p}\, d\bar{x}$$



$$KL(p\|q) \xrightarrow{q} \min \qquad q(x) = \mathcal{N}(x \mid \mu, \sigma^2)$$

$$= \int p(x) \cdot \log \frac{p}{q}\, d\bar{x}$$

$$\int p(\bar{x}) \underline{\log q(\bar{x})}\, d\bar{x} = \int p(\bar{x}) \cdot \left(-\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(x-\mu)^2 \right) dx$$

$$= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\int p(x)\left(x^2 - 2x\mu + \mu^2\right) dx = (\mathbb{E}x)^2 + \operatorname{Var} X$$

$$= -\frac{1}{2}\log\sigma^2 - \frac{\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \cdot \overbrace{\int x\, p(x)\, dx}^{\mathbb{E}_p x} - \frac{1}{2\sigma^2}\overbrace{\int x^2 p(x)\, dx}^{=} =$$

$$= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\mu^2 - 2\mu\,\mathbb{E}_p x + (\mathbb{E}_p x)^2\right) - \frac{\operatorname{Var}_p[x]}{2\sigma^2} =$$

$$= -\left(\frac{1}{2\sigma^2} \underbrace{\left(\mu - \mathbb{E}_p[x]\right)^2} + \frac{1}{2} \underbrace{\left(-\log \sigma^2 + \frac{1}{\sigma^2}\operatorname{Var}_p[x] \right)} \right) \longrightarrow \max$$

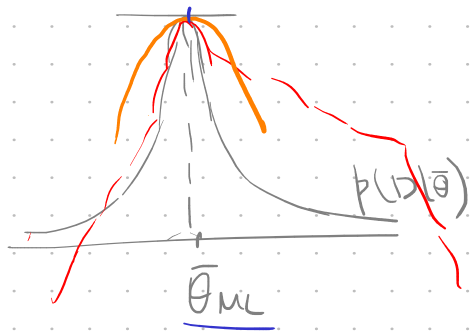
$$-\frac{1}{1/\sigma^2} + \operatorname{Var} = 0$$

moment matching
$$\begin{cases} \mu_* = \mathbb{E}_p[x] \\ \sigma^2_* = \operatorname{Var}_p[x] \end{cases}$$

$$p(D) = \int p(\bar{\theta})\, p(D|\bar{\theta})\, d\bar{\theta} = \int \underline{p(\bar{\theta})} \cdot e^{\ell(\bar{\theta})}\, d\bar{\theta}$$

$$\ell(\bar{\theta}) = \log p(D|\bar{\theta})$$

$$\ell(\bar{\theta}) \approx \ell(\bar{\theta}_{ML}) + \frac{1}{2}(\bar{\theta} - \bar{\theta}_{ML})^T \cdot H \cdot (\bar{\theta} - \bar{\theta}_{ML})$$



$$= \ell(\bar\theta_{ML}) - \frac{N}{2}(\bar\theta-\bar\theta_{ML})^T \cdot J(\bar\theta_{ML})\,(\bar\theta-\bar\theta_{ML}), \quad \text{με}$$

$$J(\bar\theta_{ML}) = -\frac{1}{N}\cdot \frac{\partial^2 \log p(D|\bar\theta)}{\partial\bar\theta\,\partial\bar\theta^T}\Bigg|_{\bar\theta=\bar\theta_{ML}}$$

$$p(\bar\theta) \approx p(\bar\theta_{ML}) + (\bar\theta-\bar\theta_{ML})^T\cdot\nabla_{\bar\theta}\, p(\bar\theta)\Big|_{\bar\theta_{ML}}$$

$$p(D) \approx \int \left(p(\bar\theta_{ML}) + \cancel{\underline{(\bar\theta-\bar\theta_{ML})^T}\, \underline{\nabla_{\bar\theta}\, p(\bar\theta)\big|_{\bar\theta_{ML}}}}\right) \cdot \underline{e^{\ell(\bar\theta_{ML}) - \frac{N}{2}(\bar\theta-\bar\theta_{ML})^T J(\bar\theta_{ML})(\bar\theta-\bar\theta_{ML})}}\, d\bar\theta =$$

$$c\cdot \bar a^T \int \underbrace{(\bar\theta-\bar\theta_{ML})}\cdot \underbrace{e^{-\frac{N}{2}(\bar\theta-\bar\theta_{ML})^T J\,(\bar\theta-\bar\theta_{ML})}}\, d\bar\theta =$$

$$= \text{const}\cdot \bar a^T\, E_{N(\bar\theta|\bar\theta_{ML}, J^{-1})}\left[\bar\theta-\bar\theta_{ML}\right] = 0$$

$$= e^{\ell(\bar\theta_{ML})}\, p(\bar\theta_{ML})\cdot \underbrace{\int e^{-\frac{N}{2}(\bar\theta-\bar\theta_{ML})^T J(\bar\theta_{ML})(\bar\theta-\bar\theta_{ML})}\, d\bar\theta}_{= (2\pi)^{d/2}\cdot\sqrt{\det(N\cdot J)^{-1}}} =$$

$$= e^{\ell(\bar\theta_{ML})}\, p(\bar\theta_{ML})\cdot (2\pi)^{\frac{d}{2}}\cdot N^{-\frac{d}{2}}\cdot (\det J)^{-1/2}$$

$$\log p(D) \approx \underbrace{\ell(\bar\theta_{ML})}_{\log p(D|\bar\theta_{ML})} + \underbrace{\overbrace{\log p(\bar\theta_{ML})}^{d\cdot(\text{const})} + \overbrace{\frac{d}{2}\log 2\pi}^{d\cdot\text{const}} - \overbrace{\frac{d}{2}\log N} - \overbrace{\frac{1}{2}\log\det J(\bar\theta_{ML})}^{\text{const}}}_{\text{Occam's factor}}$$

$$\boxed{\log p(D) \approx \log p(D \mid \bar{\theta}_M) - \frac{1}{2} \cdot d \log N}$$

BIC

Bayesian information criterion

$$\boxed{BIC(M) = -2 \log p(D \mid \bar{\theta}_M) + d \log N} \xrightarrow{M} \min$$