Exponential family

$$p(\bar{x}|\bar{\theta}) = h(\bar{x})\, e^{\bar{\eta}(\bar{\theta})^T \bar{t}(\bar{x}) - a(\bar{\theta})} = h(\bar{x})\, g(\bar{\theta})\, e^{\bar{\eta}(\bar{\theta})^T \bar{t}(\bar{x})}$$

sufficient statistics

$$p(D|\bar{\theta}) = \ldots(\bar{\theta}, \sum_n \bar{t}(\bar{x}_n))$$

natural parameter

$$p(\bar{x}|\bar{\theta}) = h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x}) - a(\bar{\theta})} = h(\bar{x})\, g(\bar{\theta})\, e^{\bar{\theta}^T \bar{t}(\bar{x})}$$

$$\mathrm{Bin}(k|n,p) = \binom{n}{k} e^{k \log p + (n-k)\log(1-p)} = \binom{n}{k} e^{k \log \frac{p}{1-p} + n\log(1-p)}$$

$-a(\bar{\theta})$

$$\theta = \log \frac{p}{1-p} \qquad n\log(1-p) = n\log(1-\sigma(\theta))$$

$$\mathcal{N}(\bar{x}|\bar{\mu},\Sigma) = e^{-\frac{1}{2}\left(\bar{x}^T\Sigma^{-1}\bar{x} - 2\bar{x}^T\Sigma^{-1}\bar{\mu} + \bar{\mu}^T\Sigma^{-1}\bar{\mu} + \log 2\pi \det\Sigma\right)}$$

$$\bar{\theta} = \begin{pmatrix} -\frac{1}{2}\mathrm{vec}(\Sigma^{-1}) \\ \Sigma^{-1}\bar{\mu} \end{pmatrix} \qquad \bar{t}(\bar{x}) = \begin{pmatrix} \mathrm{vec}(\bar{x}\bar{x}^T) \\ \bar{x} \end{pmatrix}$$

cumulant

$\boxed{\mathrm{I}}$

$$p(\bar{x}|\bar{\theta}) = h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x}) - a(\bar{\theta})}$$

$$1 = \int p(\bar{x}|\bar{\theta})\, d\bar{x} = e^{-a(\bar{\theta})} \cdot \int h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x})}\, d\bar{x}$$

$$\cancel{\nabla_{\bar{\theta}}} \quad a(\bar{\theta}) = \log \int h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x})}\, d\bar{x}$$

$$\nabla_{\bar{\theta}} a(\bar{\theta}) = \nabla_{\theta} \log \int \ldots = \frac{\nabla_{\bar{\theta}} \int h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x})}\, d\bar{x}}{\int \ldots} = \frac{\int h(\bar{x}) \cdot \nabla_{\bar{\theta}} e^{\bar{\theta}^T \bar{t}(\bar{x})}\, d\bar{x}}{e^{a(\bar{\theta})} = \int h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x})}\, d\bar{x}} =$$

$$= \frac{\int h(\bar{x}) \cdot e^{\bar{\theta}^T \bar{t}(\bar{x})} \cdot \bar{t}(\bar{x})\, d\bar{x}}{\cancel{\int \ldots d\bar{x}}\ e^{a(\bar{\theta})}} = \int \bar{t}(\bar{x}) \cdot \big(\underbrace{h(\bar{x})\, e^{\bar{\theta}^T \bar{t}(\bar{x}) - a(\bar{\theta})}}_{p(\bar{x}|\bar{\theta})}\big)\, d\bar{x}$$

$$\boxed{\nabla_{\bar\theta}\, a(\bar\theta) = \mathbb{E}_{p(\bar x|\bar\theta)}[\bar t(\bar x)]}$$

$$\boxed{\bar\mu = \mathbb{E}[\bar t(\bar x)] = \nabla_{\bar\theta}\, a(\bar\theta)}$$

- mean parametrization

$$\frac{\partial^2 a}{\partial\theta_i \partial\theta_j} = \frac{\partial\, \mathbb{E}_p[t_i(\bar x)]}{\partial\theta_j} = \frac{\partial}{\partial\theta_j}\int t_i(\bar x) h(\bar x) e^{\bar\theta^T \bar t(\bar x) - a(\bar\theta)} d\bar x$$

$$= \int t_i(\bar x) h(\bar x)\, e^{\bar\theta^T \bar t(\bar x) - a(\bar\theta)} \left(t_j(\bar x) - \frac{\partial a}{\partial\theta_j}\right) d\bar x =$$

$$= \int p(\bar x|\bar\theta) \cdot t_i(\bar x)\left(t_j(\bar x) - \mathbb{E}_p[t_j(\bar x)]\right) d\bar x =$$

$$= \mathbb{E}_{p(\bar x|\bar\theta)}[t_i(\bar x) t_j(\bar x)] - \mathbb{E}_p[t_i(\bar x)]\,\mathbb{E}_p[t_j(\bar x)]$$

$$\boxed{H(a) = \left(\frac{\partial^2 a}{\partial\theta_i \partial\theta_j}\right)_{i,j=1}^{d} = \mathrm{Var}[\bar t(\bar x)]}$$

II Max likelihood

$$p(\bar x_1 \ldots \bar x_N|\bar\theta) = \prod_n h(\bar x_n)\, e^{\bar\eta(\bar\theta)^T \bar t(\bar x_n) - a(\bar\theta)}$$

$$\log p(\bar x_1 \ldots \bar x_N|\bar\theta) = \sum_n \log h(\bar x_n) + \bar\eta(\bar\theta)^T\left(\sum_n^N \bar t(\bar x_n)\right) - N \cdot a(\bar\theta) \xrightarrow[\bar\theta]{} \max$$

Pitman – Koopman – Darmois theorem

$$\nabla_{\bar\theta}\, \bar\eta(\bar\theta)^T\left(\sum_{n=1}^{N} \bar t(\bar x_n)\right) - N \cdot \nabla_{\theta}\, a(\bar\theta) = 0$$

$$\nabla_\theta\, a(\bar\theta) = \frac{1}{N}\cdot\left(\sum t(\bar x_n)\right)^T \nabla_\theta\, \bar\eta(\bar\theta)$$

Εάν $\bar{\eta}(\bar{\theta}) = \bar{\theta}$, τότε

$$\nabla_{\bar{\theta}}\, a(\bar{\theta}) = \frac{1}{N} \cdot \sum_{n=1}^{N} \bar{t}(\bar{x}_n) = \bar{\mu}_{ML}$$

III **Bayesian inference** – conjugate prior

$$\underset{= p(\bar{\theta}|\bar{\alpha})}{p(\bar{\theta})} \times p(\bar{x}_1 \ldots \bar{x}_N | \bar{\theta}) \propto \underset{= p(\bar{\theta}|\bar{\alpha}')}{\underline{p(\bar{\theta}|\bar{x}_1 \ldots \bar{x}_N)}}$$

$$p(\bar{\theta}|\bar{\chi}, \nu) = f(\bar{\chi}, \nu) \cdot e^{\bar{\chi}^T \bar{\eta}(\bar{\theta}) - \nu \cdot a(\bar{\theta})}$$

$$\log p(\bar{\theta}|\bar{x}_1 \ldots \bar{x}_N, \bar{\chi}_0, \nu_0) = \text{Const} + \log p(\bar{\theta}|\bar{\chi}_0, \nu_0) + \log p(\bar{x}_1 \ldots \bar{x}_N|\bar{\theta})$$

$$= \text{Const} + \bar{\chi}_0^T \bar{\eta}(\bar{\theta}) - \nu_0 a(\bar{\theta}) + \sum_n \log h(\bar{x}_n) + \bar{\eta}(\bar{\theta})^T \left(\sum \bar{t}(\bar{x}_n)\right) - N a(\bar{\theta})$$

$$= \text{Const} + \left(\bar{\chi}_0 + \sum \bar{t}(\bar{x}_n)\right)^T \bar{\eta}(\bar{\theta}) - (\nu_0 + N)\, a(\bar{\theta})$$

$$p(\bar{\theta}|\bar{x}_1 \ldots \bar{x}_N, \bar{\chi}_0, \nu_0) = p(\bar{\theta}|\bar{\chi}_N, \nu_N), \text{ όπου}$$

$$\begin{cases} \nu_N = \nu_0 + N \\ \bar{\chi}_N = \bar{\chi}_0 + \sum_{n=1}^{N} \bar{t}(\bar{x}_n) \end{cases}$$

$$\mathcal{N}(x|\mu, \tau) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \cdot e^{-\frac{\tau}{2}x^2 + \tau\mu x - \frac{\mu^2}{2}\tau} = \frac{1}{\sqrt{2\pi}} e^{\begin{pmatrix} -\tau/2 \\ \tau\mu \end{pmatrix}^T \begin{pmatrix} x^2 \\ x \end{pmatrix} - \frac{\mu^2 \tau}{2} + \frac{1}{2}\log\tau}$$

$$p(\bar{\theta}|\bar{\chi}_0, \nu_0) = f(\bar{\chi}_0, \nu_0) \cdot e^{\bar{\chi}_0^T \cdot \begin{pmatrix} -\tau/2 \\ \tau\mu \end{pmatrix} - \nu_0\left(\frac{\mu^2\tau}{2} - \frac{1}{2}\log\tau\right)} =$$

$$= f(\bar{\chi}_0, \nu_0) \cdot e^{-\chi_0 \tau/2 + \chi_1 \tau \mu - \nu_0 \frac{\mu^2 \tau}{2} + \frac{\nu_0}{2}\log\tau} =$$

$$= \underline{f(\bar{\chi}_0, \nu_0)}\, e^{-\frac{\nu_0 \tau}{2}\left(\mu - \frac{\chi_1}{\nu_0}\right)^2 + \frac{\chi_1^2}{2\nu_0}\tau - \frac{\chi_0 \tau}{2} + \frac{\nu_0}{2}\log \tau} =$$

$$= f(\bar{\chi}_0, \nu_0) \cdot \underbrace{e^{-\frac{\nu_0 \tau}{2}\left(\mu - \frac{\chi_1}{\nu_0}\right)^2}}_{\mathcal{N}\left(\mu \,\middle|\, \frac{\chi_1}{\nu_0},\ \nu_0 \tau\right)} \cdot \underbrace{\tau^{\frac{\nu_0}{2}} \cdot e^{-\left(\frac{\chi_0}{2} - \frac{\chi_1^2}{2\nu_0}\right)\tau}}_{\mathrm{Gam}\left(\tau \,\middle|\, \frac{\nu_0}{2},\ \frac{\chi_0}{2} - \frac{\chi_1^2}{2\nu_0}\right)}$$

$$\mathcal{N}\left(\mu \,\middle|\, \underbrace{\frac{\chi_1}{\nu_0}}_{\mu_0},\ \underbrace{\nu_0}_{\lambda_0}\tau\right) \times \mathrm{Gam}\left(\tau \,\middle|\, \underbrace{\frac{\nu_0}{2}}_{\alpha_0},\ \underbrace{\frac{\chi_0}{2} - \frac{\chi_1^2}{2\nu_0}}_{\beta_0}\right)$$

$\mathcal{N}(x \mid \tau)$ $\qquad$ $\mathrm{Gam}(\tau \mid \alpha_0, \beta_0)$

$\mathcal{N}(\bar{x} \mid \Sigma^{-1} = \Lambda)$ $\qquad$ Wishart distribution

$$W(\Lambda \mid V, n, d) = \frac{1}{\text{Const} \cdot (\det V)^{n/2}} \cdot (\det \Lambda)^{\frac{n-d-1}{2}} \cdot e^{-\frac{1}{2}\mathrm{Tr}(V^{-1}\Lambda)}$$

$\mathcal{N}(\bar{x} \mid \bar{\mu}, \Lambda)$ $\qquad$ $W(\Lambda) \cdot \mathcal{N}(\bar{\mu} \mid \bar{\mu}_0, \ldots \Lambda)$

$\boxed{\text{IV}}$ Predictive distribution

$$p(\bar{x} \mid D) = \int p(\bar{x} \mid \bar{\theta})\, \underline{p(\bar{\theta} \mid D)}\, d\bar{\theta} = \int p(\bar{x} \mid \bar{\theta}) \left(\frac{p(\bar{\theta} \mid \bar{\chi}_0, \nu_0) \cdot p(D \mid \bar{\theta})}{p(D \mid \bar{\chi}_0, \nu_0)} \right) d\bar{\theta}$$

$$p(\bar{\theta} \mid \bar{\chi}_0, \nu_0) \times p(D \mid \bar{\theta}) \propto p(\bar{\theta} \mid \bar{\chi}_N, \nu_N)$$

$$= \int p(\bar{x} \mid \bar{\theta}) \cdot p(\bar{\theta} \mid \bar{\chi}_N, \nu_N)\, d\bar{\theta} = \int h(\bar{x})\, e^{T(\bar{x})^T \eta(\bar{\theta}) - a(\bar{\theta})} \times$$

$$\times f(\bar{\chi}_N, \nu_N) \cdot e^{\bar{\chi}_N^T \eta(\bar{\theta}) - \nu_N \cdot a(\bar{\theta})}\, d\bar{\theta} =$$

$$= f(\bar{\chi}_N, \nu_N) \cdot \int h(\bar{x}) \int e^{(\bar{\chi}_N + T(\bar{x}))^T \eta(\bar{\theta}) - (\nu_N + 1)\, a(\bar{\theta})}\, d\bar{\theta} =$$

$$= h(\bar{x}) \cdot \frac{f(\bar{\chi}_N, \nu_N)}{f(\bar{\chi}_N + \bar{t}(\bar{x}), \nu_N + 1)}$$

$$p(\bar{x} \mid D) = h(\bar{x}) \frac{f(\bar{\chi}_N, \nu_N)}{f(\bar{\chi}_N + \bar{t}(\bar{x}), \nu_N + 1)}$$

$$p(x^{=1} \mid \theta) = p = \sigma(\theta)$$

$$\theta = \log \frac{p}{1-p} \qquad p = (1-p) e^{\theta}$$

$$p = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}}$$

$$\mathrm{Bin}(k \mid n, p) = \binom{n}{k} e^{k \cdot \theta - (-n \log(1-\sigma(\theta)))}$$

$$= \log \frac{e^{-\theta}}{1+e^{-\theta}} = -\theta - \log(1+e^{-\theta})$$

$$p(\theta) = f(-) \cdot e^{x \cdot \theta + \nu \cdot n \log(1-\sigma(\theta))}$$

Generalized linear models / GLM

LR: $(\bar{w}^T \bar{x}) \sim y$, $h(a) = a$

log-regr.: $\sigma(\bar{w}^T \bar{x}) \sim y$, $h(a) = \sigma(a)$

probit: $h(a) = \Phi(a)$

$$\hat{y} = h(\bar{w}^T \bar{x})$$

$$c = \bar{w}^T \bar{x} \qquad h(c) = \mu \overset{!}{=} \mathbb{E}_{x}[y]$$

$\bar{w}^T \bar{x} = c = g(\mu)$, $\mu = g^{-1}(c)$ $\quad$ g — link function

LR: $g = id$ $\quad$ LogR: $g(\mu) = \sigma^{-1}(\mu) = \log \frac{\mu}{1-\mu}$

$p(y|\bar{x},\bar{w})$: $\mu = g^{-1}(\bar{w}^T\bar{x})$ — sufficient statistic

---

$$p(y|\theta,\sigma^2) = h(y,\sigma^2)\cdot e^{\frac{y\theta - a(\theta)}{\sigma^2}}$$

$\theta = \bar{w}^T\bar{x}$, $a(\theta,\sigma^2)$, $t(y)$, h

overdispersed exponential family

$$E\left[\frac{y}{\sigma^2}\right] = \frac{\partial\left(\frac{1}{\sigma^2}a(\theta)\right)}{\partial\theta}$$

$\theta = \bar{w}^T\bar{x}$

$$E[y] = \frac{\partial a}{\partial\theta}$$

$$Var\left[\frac{y}{\sigma^2}\right] = \frac{1}{\sigma^4}\cdot Var[y] = \frac{\partial^2\left(\frac{1}{\sigma^2}a(\theta)\right)}{\partial\theta^2}$$

$$Var[y] = \sigma^2\cdot\frac{\partial^2 a(\theta)}{\partial\theta^2}$$

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y-\mu)^2} =$$

$\mu = \bar{w}^T\bar{x}$

$$= e^{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}$$

$g = g^{-1} = id$ $\quad \theta = \mu = \bar{w}^T\bar{x}$, $a(\theta) = a(\mu) = \frac{\mu^2}{2} = \frac{\theta^2}{2}$

$$h(y,\sigma^2) = e^{-\left(y^2/2\sigma^2 + \frac{1}{2}\log(2\pi\sigma^2)\right)}$$

$$p(y|\mu) = \mu^y(1-\mu)^{1-y} = e^{y\log\frac{\mu}{1-\mu} + \log(1-\mu)}$$

$\theta = \log\frac{\mu}{1-\mu}$, $\sigma^2 = 1$, $a(\theta) = -\log(1-\mu)$, $h(y,\sigma^2) = 1$

$$\mu = \sigma(\theta) = \frac{1}{1+e^{-\theta}}$$

$$= -\log\left(1 - \frac{1}{1+e^{-\theta}}\right) = -\log\frac{e^{-\theta}}{1+e^{-\theta}} = \log(1+e^{\theta})$$

$$g(\mu) = \log \frac{\mu}{1-\mu}, \quad g^{-1}(\theta) = \sigma(\theta)$$

$$\frac{\partial a}{\partial \theta} = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} = \sigma(\theta) = g^{-1}(\theta) = \mu$$

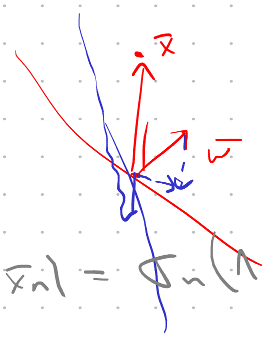
$$\log p(y|\bar{x}, \bar{w}, \sigma^2) = \underline{\log h(y, \sigma^2)} + \frac{y}{\sigma^2} \bar{w}^T \bar{x} - \frac{1}{\sigma^2} a(\bar{w}^T \bar{x})$$

$$\log p(D|X, \bar{w}, \sigma^2) = Const + \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(y_n \cdot \bar{w}^T \bar{x}_n - a(\bar{w}^T \bar{x}_n) \right) \xrightarrow{\bar{w}} \max$$

$$\nabla_{\bar{w}} \log p = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(y_n \cdot \bar{x}_n - a'(\bar{w}^T \bar{x}_n) \cdot \bar{x}_n \right) =$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(\underline{y_n} \underbrace{- a'(\bar{w}^T \bar{x}_n)}_{= \mu_n} \right) \cdot \underline{\bar{x}_n}$$

$\sigma(\bar{x}^T \bar{w})$



$$\nabla_{\bar{w}} \log p = \frac{1}{\sigma^2} \sum_{n} (y_n - \mu_n) \bar{x}_n$$

$$\mu'(\cdot) = \sigma'(\bar{w}^T \bar{x}_n) = \sigma_n (1 - \sigma_n)$$

$$H(\log p) = -\frac{1}{\sigma^2} \sum_{n=1}^{N} \underline{a''(\bar{w}^T \bar{x}_n)} \cdot \bar{x}_n \bar{x}_n^T =$$

$$= -\frac{1}{\sigma^2} X^T \cdot \begin{pmatrix} \cdot & & \\ & \ddots & \\ & & \cdot \end{pmatrix} X$$

$p(x|\lambda) = \frac{1}{x!} \lambda^x e^{-\lambda}$

Var $= \lambda$

<u>Poisson regression</u>

y

$$p(y|\mu) = \frac{1}{y!} \mu^y \cdot e^{-\mu}$$

$$\log p(y|\mu) = y \underbrace{\log \mu}_{= \theta} - \mu - \log(y!)$$

$\sigma^2 = 1$, $\theta = \log \mu$, $\log \mu = \theta = \overline{w}^T \overline{x}$, $a(\theta) = e^{\theta} = \mu$

$$\log p(y \mid \overline{x}, \overline{w}) = y \cdot \overline{w}^T \overline{x} - e^{\overline{w}^T \overline{x}} - \log(y!)$$

Negative Binomial

$$\text{NegBinom}(k \mid r, p) = \binom{k+r-1}{r-1} (1-p)^k p^r$$

$NB(k \mid 1, p)$

$$\mu = \frac{r(1-p)}{p} \; ; \; p = \frac{r}{\mu + r}$$

$$p(y \mid r, p) = \binom{y+r-1}{r-1} (1-p)^y p^r$$

$$\text{Var}(y \mid \mu, r) = \mu + \frac{\mu^2}{r} = r \cdot \frac{1-p}{p} + r \cdot \frac{(1-p)^2}{p^2}$$

$$p(y \mid r, p) = \binom{y+r-1}{r-1} \left(\frac{r}{\mu+r}\right)^r \cdot \left(\frac{\mu}{\mu+r}\right)^y =$$

$$\boxed{\underbrace{\log \mu}_{=\theta} \sim \overline{w}^T x}$$

$$= (\ldots) \cdot e^{r \log \frac{r}{\mu+r} + y \log \frac{\mu}{\mu+r}} =$$

$$= (\_\_) \, e$$

$p(D \mid \overline{\theta}) \longrightarrow \max$ $\quad \overline{\theta}_{MLE}$ MLE

$p(\overline{\theta})\, p(D \mid \overline{\theta}) \longrightarrow \max$ $\quad \overline{\theta}_{MAP}$

$\propto p(\overline{\theta} \mid D)$

$\overline{\theta}^{(r)} \sim p(\overline{\theta} \mid D)$

$p(\overline{w} \mid D) \sim \overline{w}^{(1)} \_\_ \overline{w}^{(k)}$

predictive distrib $\longrightarrow$ $p(\overline{x} \mid D) \approx \frac{1}{R} \sum_r p(\overline{x} \mid \overline{\theta}^{(r)})$

$$y = \left(g(\bar{x}) + \varepsilon_1\right)^2 \cdot \left(h(\bar{x}) + \varepsilon_2\right)$$

$$\log y = 2\log\left(g(x) + \varepsilon_1\right) + \log\left(h(x) + \varepsilon_2\right)$$