$\boxed{\text{Tutorial 2}}$

CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Problem 2-6: MLE for m.v. Gaussian

$$D = \{x_i\}_{i=1}^N$$

$$\log p(D) = \sum_i \log p(x_i)$$

$$= \sum_i \left[\log \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\|x_i-\mu\|^2_\Sigma}\right]$$

$$= \sum_i \left[\underbrace{-\frac{d}{2}\log 2\pi}_{x} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\|x_i-\mu\|^2_\Sigma\right]$$

$$= -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_i \|x_i-\mu\|^2_\Sigma + \text{const}$$

a) $\hat{\mu} = \arg\max_\mu -\frac{1}{2}\sum_i \|x_i-\mu\|^2_\Sigma$

$$= \arg\max_\mu \underbrace{-\frac{1}{2}\sum_i (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}_{\phi}$$

$$\frac{\partial}{\partial x} x^TAx = Ax + A^Tx = 2Ax \quad (A \text{ is symmetric})$$

$$\frac{\partial}{\partial x} ax^2 = 2ax$$

$$\frac{\partial \phi}{\partial \mu} = -\frac{1}{2}\sum_i 2\Sigma^{-1}(x_i-\mu)(-1) = 0$$

$$\sum_i (x_i-\mu) = 0$$

$$\sum_i x_i - N\mu = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N}\sum_i x_i}$$

b) $\hat{\Sigma} = \arg\max_\Sigma -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_i (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$

$$\frac{\partial}{\partial \Sigma} -\frac{N}{2}\log|\Sigma|$$

$$-\frac{N}{2}\Sigma^{-T} = -\frac{N}{2}\Sigma^{-1}$$

$$\frac{\partial}{\partial X}\log|X| = X^{-T}$$

$$\frac{\partial}{\partial x}\log x = \frac{1}{x}$$

$$\frac{\partial}{\partial \Sigma}\sum_i \underbrace{(x_i-\mu)^T}_{a}\underbrace{\Sigma^{-1}(x_i-\mu)}_{b}$$

$$\underbrace{a^Tb}_{\mathbb{R}} = tr(a^Tb) = tr(ba^T)$$

$$\frac{\partial}{\partial \Sigma}\sum_i tr(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T)$$

$$\frac{\partial}{\partial \Sigma} tr(\underbrace{\Sigma^{-1}}_{X^{-1}}\underbrace{\sum_i (x_i-\mu)(x_i-\mu)^T}_{A})$$

$$\frac{\partial}{\partial X} tr(AX^{-1}) = \frac{\partial}{\partial X} tr(X^{-1}A) = -(X^{-T}A^TX^{-T})$$

$$= -\Sigma^{-1}\sum_i (x_i-\mu)(x_i-\mu)^T\Sigma^{-1}$$

$$\frac{\partial}{\partial x}\frac{a}{x} = \frac{-a}{x^2}$$

$$\Rightarrow \frac{\partial}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left[\sum_i (x_i-\mu)(x_i-\mu)^T\right]\Sigma^{-1} = 0$$

$\Sigma \times \quad \times\Sigma \times 2$

$$-N\Sigma + \sum_i (x_i-\mu)(x_i-\mu)^T = 0$$

$$\boxed{\Rightarrow \hat{\Sigma} = \frac{1}{N}\sum_i (x_i-\mu)(x_i-\mu)^T}$$

alternatively

$$\frac{\partial}{\partial \Sigma^{-1}}\log|\Sigma| = \frac{\partial}{\partial \Sigma^{-1}} -\log|\Sigma^{-1}| = -\Sigma$$

$$\frac{\partial}{\partial \Sigma^{-1}} tr(\Sigma^{-1}A) = A$$

$$+\frac{N}{\cancel{2}}(+\Sigma) - \frac{1}{\cancel{2}}A = 0$$

$$\Sigma = \frac{1}{N}A$$

PS 2.8 LS regression

$$f(x) = \underbrace{\phi(x)}_{\text{features}}{}^T \underbrace{\theta}_{\text{parameters}}$$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \quad \text{(polynomials)} \qquad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_k \end{bmatrix}$$

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Assume a dataset: $D = \{(x_i, y_i)\}_{i=1}^N$

a) Least squares

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \Phi = \begin{bmatrix} | & & | \\ \phi(x_1) & \dots & \phi(x_N) \\ | & & | \end{bmatrix}$$

minimize squared error: $$\sum_i (y_i - \underbrace{\phi(x_i)^T\theta}_{f(x_i)})^2 = \underbrace{\begin{bmatrix} y_1 - \phi(x_1)^T\theta \\ \vdots \\ y_N - \phi(x_N)^T\theta \end{bmatrix}^T}_{y - \Phi^T\theta} \begin{bmatrix} y_1 - \phi(x_1)^T\theta \\ \vdots \\ y_N - \phi(x_N)^T\theta \end{bmatrix}$$

$$= \|y - \Phi^T\theta\|^2$$

$$= (y - \Phi^T\theta)^T (y - \Phi^T\theta)$$

$$= y^Ty - \underbrace{(\Phi^T\theta)^T y}_{\underbrace{-\theta^T\Phi y}_{\text{scalar}}} - \underbrace{y^T\Phi^T\theta}_{\text{scalar}} + \theta^T\Phi\Phi^T\theta$$

$$\underbrace{y^T\Phi^T}_{\text{vector}}\underbrace{\theta}_{\text{vector}} = \text{number}$$

$$= \underbrace{y^Ty}_{x} - \underbrace{2y^T\Phi^T\theta}_{a^T\theta} + \underbrace{\theta^T\Phi\Phi^T\theta}_{\theta^TA\theta}$$

$$\frac{\partial}{\partial\theta} = a \qquad \frac{\partial}{\partial\theta} = 2A\theta$$

minimize: $$\frac{\partial}{\partial\theta}(\cdot) = -\cancel{2}\Phi y + \cancel{2}\Phi\Phi^T\theta = 0$$

$$(\Phi\Phi^T)\theta = \Phi y$$

$$\Rightarrow \boxed{\hat{\theta} = (\Phi\Phi^T)^{-1}\Phi y}$$

$$\Phi = \underbrace{\begin{bmatrix} \phi(x_1) & \dots & \phi(x_N) \end{bmatrix}}_{N} \; \Big\updownarrow K+1$$

Using Gaussian noise:

b) $$\ell = \sum_i \log p(y_i \mid x_i, \theta)$$

$$= \sum_i \log N(y_i \mid f(x_i), \sigma^2)$$

$$= \sum_i -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \underline{\frac{1}{2\sigma^2}(y_i - f(x_i))^2}$$

MLE: $$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \ell$$

$$= \underset{\theta}{\operatorname{argmax}} \underbrace{-\frac{1}{2\sigma^2}} \sum_i (y_i - f(x_i))^2$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_i (y_i - f(x_i))^2 \leftarrow \text{LS formulation}$$

P2-10 $\Rightarrow$ Robust Regression.