Hw2writeup

Thursday, October 27, 2016 9:49 PM

(i) $y^i \in \mathbb{R}^n$ $x^i \in \mathbb{R}^n$

$Y = [y^1\ y^2\ y^3 \dots y^m]$ $X = [x^1\ x^2\ x^3 \dots x^m]$

$m > n$ and XX^T is invertible.

$$L_{ls} = \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) \quad -(1)$$

$$(Y - AX)^T (Y - AX)_{ij} = (y^i - Ax^i)^T (y^j - Ax^j)$$

$\therefore$ diagnol element of $(Y-AX)^T(Y-AX)$

$$(Y-AX)^T(Y-AX)_{ii} = (y^i - Ax^i)^T (y^i - Ax^i)$$

$$\Rightarrow L_{ls} = \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) = Tr[(Y-AX)^T(Y-AX)]$$

$$\therefore A_{LS} = \arg\min_A Tr[(Y-AX)^T(Y-AX)]$$

$$\frac{\partial L_{ls}}{\partial A} = 0 \quad (\text{for minimising w.r.t } A)$$

$$\frac{\partial L_{ls}}{\partial A} = \frac{\partial (Tr[(Y-AX)^T(Y-AX)])}{\partial A} = 0$$

$$\Rightarrow \frac{\partial (Tr[Y^TY + X^TA^TAX - 2Y^TAX])}{\partial A} = 0$$

$$\Rightarrow \underset{(1)}{\frac{\partial Tr(Y^TY)}{\partial A}} + \underset{(2)}{\frac{\partial Tr(X^TA^TAX)}{\partial A}} - 2\underset{(3)}{\frac{\partial (Tr(Y^TAX))}{\partial A}} = 0$$

as $(Tr(A+B) = Tr(A) + Tr(B)$ if A and B both $m \times m$ matrix).

$$\Rightarrow \underset{(1)}{0} + \underset{(2)}{AXX^T + AXX^T} - \underset{(3)}{2YX^T} = 0$$

(using derivatives of traces).

$$\cancel{2} AXX^T - \cancel{2} YX^T = 0$$

$$\therefore \boxed{A_{LS} = (YX^T)(XX^T)^{-1}}$$

(ii) $L_r = \lambda \|A\|_F^2 + \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i)$

$$L_r = \lambda Tr(A^TA) + \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i)$$

from part (i) we know

$$L_r = \lambda \, Tr(A^T A) + Tr[(Y-AX)^T (Y-AX)]$$

$$A_r = \arg\min_A L_r$$

$$A_r = \arg\min_A \left(\lambda Tr(A^T A) + Tr[(Y-AX)^T (Y-AX)]\right)$$

$\therefore \dfrac{\partial L_r}{\partial A} = 0$ (to find minimum value and corresponding A)

$$\frac{\partial L_r}{\partial A} = \frac{\partial[\lambda Tr(A^T A) + Tr[(Y-AX)^T(Y-AX)]]}{\partial A} = 0$$

using the derivatives of the traces.

$$2\lambda A + 2AXX^T - 2YX^T = 0$$

$$A(XX^T + \lambda I) = YX^T$$

$$A = (YX^T)(XX^T + \lambda I)^{-1}$$

$$\therefore A_r = (YX^T)(XX^T + \lambda I)^{-1}$$

(iii) $\varepsilon_i = y^i - Ax^i \sim N(0, \sigma^2 I)$

$f(\varepsilon_i)$ = p.d.f of ε_i

$$f(\varepsilon_i) = (2\pi)^{-n/2} |\sigma^2 I|^{-\frac{1}{2}} \exp\left(\frac{-1}{2\sigma^2} (\varepsilon_i - 0)^T (\varepsilon_i - 0)\right)$$

likelihood fn $\ell(\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots \varepsilon_m | A, \sigma^2)$

$$= f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m | A, \sigma^2)$$

$$= \prod_{i=1}^{m} f(\varepsilon_i | A, \sigma^2) \quad [\text{assuming } y^i - Ax^i \text{ independent}]$$

$$\ell(\varepsilon_1 \ldots \varepsilon_m | A, \sigma^2) = \prod_{i=1}^{m} (2\pi)^{\frac{-n}{2}} (\sigma^2)^{\frac{-n}{2}} \exp\left(\frac{-1}{2\sigma^2} (y^i - Ax^i)^T (y^i - Ax^i)\right)$$

$$\Rightarrow 2\pi^{\frac{-nm}{2}} \sigma^{-mn} \exp\left(\sum_{i=1}^{m} \frac{-1}{2\sigma^2} (y^i - Ax^i)^T (y^i - Ax^i)\right)$$

$$\Rightarrow \ell(\varepsilon_1 \ldots \varepsilon_m | A, \sigma^2) = (2\pi)^{\frac{-mn}{2}} \sigma^{-mn} \exp\left(\frac{-1}{2\sigma^2} \sum (y^i - Ax^i)^T (y^i - Ax^i)\right)$$

$$\log[\ell(\varepsilon_1, \varepsilon_2 \dots \varepsilon_m | A, \sigma^2)]$$

$$= \log(2\pi)^{\frac{-mn}{2}} + \log(\sigma^{-mn}) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)$$

$$A_{MLE} = \arg\max_{A} \log[\ell(\varepsilon_1, \varepsilon_2, \dots \varepsilon_m | A, \sigma^2)]$$

$$A_{MLE} = \arg\max_{A} \Big[\underbrace{\log(2\pi)^{\frac{-mn}{2}} + \log \sigma^{-mn}}_{\text{does not depend on } A.} - \frac{1}{2\sigma^2}\Big[\sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)\Big]\Big]$$

$$\therefore A_{MLE} = \arg\max_{A} \Big[\underbrace{\boxed{\frac{-1}{2\sigma^2}}}_{= -ve} \sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)\Big]$$

$$\Rightarrow A_{MLE} = \arg\min_{A} \Big[\underbrace{\sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)}_{\text{Same as } L_{ls}}\Big]$$

$$\frac{\partial\left(\sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)\right)}{\partial A} = 0$$

$$\Rightarrow \sum_{i=1}^{m}\frac{\partial\left(y^{iT}y^i + x^{iT}A^TAx^i - 2x^{iT}A^Ty^i\right)}{\partial A} = 0$$

$$\Rightarrow \sum_{i=1}^{m}\left(0 + 2Ax^ix^{iT} - 2y^ix^{iT}\right) = 0$$

$$\Rightarrow A\left(\sum_{i=1}^{m}x^ix^{iT}\right) = \sum_{i=1}^{m}y^ix^{iT}$$

$$\sum_{i=1}^{m}x^ix^{iT} = XX^T$$

$$\sum_{i=1}^{m}y^ix^{iT} = YX^T$$

$$\therefore A_{MLE} = (YX^T)(XX^T)^{-1}$$

(iv) $\varepsilon_i \sim N(0, \sigma^2 I)$

where, $\varepsilon_i = y^i - Ax^i$

and $A \sim MN(M, \lambda^{-\frac{1}{2}}I, \lambda^{-\frac{1}{2}}I)$

$\sim MN(\ldots, \ldots I, \ldots I)$

$\pi(A) =$ Prior distribution

$f(A \mid \varepsilon_1, \varepsilon_2 \ldots \varepsilon_m) =$ Posterior distribution

$$A_{MAP} = \arg\max_{A} f(A \mid \varepsilon_1, \varepsilon_2, \ldots \varepsilon_m)$$

$$f(A \mid \varepsilon_1, \varepsilon_2 \ldots \varepsilon_m) = \frac{f(\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m, A)}{f(\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m)}$$

where $f(\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m, A)$ is joint probability distribution

using Bayes Rule.

$$f(A \mid \varepsilon_1 \ldots \varepsilon_m) = \frac{f(\varepsilon_1 \ldots \varepsilon_m \mid A)\, \pi(A)}{f(\varepsilon_1, \ldots, \varepsilon_m)}$$

$f(\varepsilon_1, \ldots, \varepsilon_m \mid A)$ is conditional joint dist^n of $\varepsilon_1, \ldots, \varepsilon_m$ on A.

$\pi(A)$ is prior dist^n of A

$f(\varepsilon_1 \ldots \varepsilon_m)$ is joint dist^n of $\varepsilon_1, \ldots \varepsilon_m$.

now

$$A_{MAP} = \arg\max_{A} \left(\frac{f(\varepsilon_1 \ldots \varepsilon_m \mid A)\, \pi(A)}{\underbrace{f(\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m)}_{\text{Independent of } A.}} \right)$$

$$\therefore A_{MAP} = \arg\max_{A} \left[f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m \mid A)\, \pi(A) \right]$$

$f(\varepsilon_1, \varepsilon_2, \ldots \varepsilon_m \mid A)\, \pi(A)$

$= f(\varepsilon_1 \mid A) f(\varepsilon_2 \mid A) \ldots f(\varepsilon_m \mid A)\, \pi(A)$ (as $\varepsilon_i = y^i - Ax^i$ and assuming independent from each other)

$$= \prod_{i=1}^{m} f(\varepsilon_i \mid A)\, \pi(A)$$

$$\prod_{i=1}^{m} (2\pi)^{-\frac{n}{2}} |\sigma^2 I|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \varepsilon_i^T \varepsilon_i \right\} \pi(A)$$

$$= (2\pi)^{-\frac{mn}{2}} \sigma^{-nm} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} \varepsilon_i^T \varepsilon_i \right\} (2\pi)^{-\frac{n^2}{2}} \exp\left\{ -\frac{1}{2} Tr(\lambda (A-M)^T (A-M)) \right\}$$

$$= \underbrace{(2\pi)^{-\frac{mn}{2} - \frac{n^2}{2}} \sigma^{-nm}}_{\text{independent of } A} \exp\left\{ \underbrace{\frac{-1}{2\sigma^2}}_{<0} \left[\sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) + Tr(\lambda\sigma^2 (A-M)^T (A-M)) \right] \right\}$$

So maximizing above expression value w.r.t. A is same as follows:

$$A_{MAP} = \arg\max_A \left(p(\varepsilon_1, \varepsilon_2, \dots \varepsilon_m | A)\, \pi(A) \right)$$

$$= \arg\max_A \text{Exp}\left\{ \frac{-1}{2\sigma^2} \left[\sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) + Tr\left(d\sigma^2 (A-M)^T (A-M)\right) \right] \right\}$$

$$A_{MAP} = \arg\min_A \text{Exp}\left\{ \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) + Tr\left(d\sigma^2 (A-M)^T (A-M)\right) \right\}$$

Since Exp() is increasing function,

$$A_{MAP} = \arg\min_A \Big(\underbrace{\sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) + Tr\left(d\sigma^2 (A-M)^T (A-M)\right)}_{f(A)} \Big)$$

$$f(A) = \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) + Tr\left(d\sigma^2 \left[A^T A + M^T M - 2A^T M\right]\right)$$

$$f(A) = Tr\left((Y - AX)^T (Y - AX) \right) + Tr\left(d\sigma^2 (A^T A - 2A^T M + M^T M) \right)$$

$\therefore \dfrac{\partial f(A)}{\partial A} = 0$ for finding A_{MAP} (minimising w.r.t A)

$$0 + 2AXX^T - 2YX^T + d\sigma^2 (2A - 2M + 0) = 0$$

$$AXX^T + d\sigma^2 A - YX^T - d\sigma^2 M = 0$$

$$A(XX^T + d\sigma^2 I) = (YX^T + d\sigma^2 M)$$

$$\Rightarrow \quad A_{MAP} = (YX^T + d\sigma^2 M)(XX^T + d\sigma^2 I)^{-1}$$

if M is zero matrix

$$A_{MAP} = (YX^T)(XX^T + d\sigma^2 I)^{-1}$$

(v) Expression derived is same for both i & ii

$$A = (YX^T)(XX^T)^{-1}$$

as both the estimates finally are derived from minimising least square error

$$\sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i)$$

where $\varepsilon_i = y^i - Ax^i$

Expression obtained in (iii) and (iv) are similar

$$A_R = (Y X^T)(X X^T + \lambda I)^{-1}$$

$$A_{MAP} = (Y X^T)(X X^T + \lambda \sigma^2 I)^{-1}$$ (if M is zero matrix)

both have a regularizer term λ or $\lambda\sigma^2$

if $A \sim MN\left(0, \left(\frac{\lambda}{\sigma^2}\right)^{-\frac{1}{2}} I, \left(\frac{\lambda}{\sigma^2}\right)^{-\frac{1}{2}} I\right)$

then both are same

or $\lambda = \lambda \sigma^2$

where $A \sim MN(0, \lambda^{-\frac{1}{2}} I, \lambda^{-\frac{1}{2}} I)$

then also both same.

Since both adding $\lambda \|A\|_F^2$ in (iii) part

or using $A \sim MN(0, \lambda^{-\frac{1}{2}} I, \lambda^{-\frac{1}{2}} I)$ in (iv) part try to regularize the the expression L_λ in (iii) or posterior in (iv) part.

both get same estimate of A with regularizer added to $X X^T$.