ORF524 - Problem Set 2

Bachir EL KHADIR

October 18, 2015

Question 1

$$\begin{aligned}
\mathbb{P}^{\theta}(X = a) &= \mathbb{P}^{\theta}(X = a, T(X) = T(a)) && (1)\\
&= \mathbb{P}^{\theta}(X = a|T(X) = T(a))\mathbb{P}^{\theta}(T(X) = T(a)) && (2)\\
&= \mathbb{P}(X = a|T(X) = T(a))\mathbb{P}^{\theta}(T(X) = T(a)) \qquad \text{By definition of sufficiency} && (3)\\
&= \mathbb{P}(X' = a|T(X) = T(a))\mathbb{P}^{\theta}(T(X) = T(a)) && (4)\\
&= \mathbb{P}^{\theta}(X' = a) && (5)
\end{aligned}$$

Question 2

Let's first note that:

$$l(\theta) = \frac{1}{\int_{\mathbb{R}^d} h(x) e^{\alpha(\theta)^T T(x)} \mathrm{dx}} = l(\alpha(\theta))$$

As a result, f^{θ} is determined entirely by $\alpha(\theta)$, we can then denote it $f_{\alpha(\theta)}$ As a result

$$P = \{f_{\alpha} | \alpha \in \alpha(\Theta)\}$$

Question 3

Since T a sufficient statistics for P_{θ}, it is also sufficient for P'_{θ}, because by definition, for all $P^{\theta} \in \mathcal{P}'^{\theta} \subseteq \mathcal{P}^{\theta}$, conditioning on the value of T makes the distribution independent of θ.

If there exist a sufficient statistics T' for P^{θ}, T is also sufficient for P'^{θ}. Since T is minimal for P'_{θ}, there exist a function ϕ such that $T = \phi(T')$, and thus T is minimal for P_{θ}.

Question 4

$$\begin{aligned}
\mathcal{N}^n_{\mu,\mu}(x) &= \frac{1}{(\sqrt{2\pi\mu})^n} e^{-\sum_i \frac{(x_i-\mu)^2}{2\mu}} && (6)\\
&= \frac{1}{(\sqrt{2\pi\mu})^n} e^{-\frac{\sum_i x_i^2}{2\mu} - \sum x_i - n\frac{\mu}{2}} && (7)\\
&= \frac{1}{(\sqrt{2\pi\mu})^n} e^{-\frac{\sum_i x_i^2}{2\mu} - n\frac{\mu}{2}} e^{-\sum x_i} && (8)\\
&= g_{\mu}(\sum x_i^2) f(x) && (9)
\end{aligned}$$

$$T(x) = \sum x_i^2$$

In the case $n = 1$, $x^2 = f(x)$, so x^2 is sufficient. But it is not minimal because x cannot be written as a function of x^2

Question 5

Let A be the constant of normalization.

$$\begin{aligned}
f_\theta^n(X) &= f^n_{(\Sigma,\mu)}(X1,..,X_n)\\
&= A\exp\{-\sum_i \frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\}\\
&= A\exp\{-\frac{1}{2}\sum_i (X_i-\hat\mu+\hat\mu-\mu)^T\Sigma^{-1}(X_i-\hat\mu+\hat\mu-\mu)\}\\
&= A\exp\{-\frac{1}{2}\sum_i \left((X_i-\hat\mu)^T\Sigma^{-1}(X_i-\hat\mu)+2(X_i-\hat\mu)^T\Sigma^{-1}(\hat\mu-\mu)+(\hat\mu-\mu)^T\Sigma^{-1}(\hat\mu-\mu)\right)\\
&= A\exp\{-\frac{1}{2}\sum_i \mathrm{Tr}\left((X_i-\hat\mu)^T\Sigma^{-1}(X_i-\hat\mu)+(\hat\mu-\mu)^T\Sigma^{-1}(\hat\mu-\mu)\right)\}\\
&= A\exp\{-\frac{1}{2}\sum_i \mathrm{Tr}\left(\Sigma^{-1}(X_i-\hat\mu)(X_i-\hat\mu)^T+\Sigma^{-1}(\hat\mu-\mu)(\hat\mu-\mu)^T\right)\}\\
&= A\exp\{-\frac{n}{2}\mathrm{Tr}\Sigma^{-1}\left(\hat\Sigma+(\hat\mu-\mu)(\hat\mu-\mu)^T\right)\}
\end{aligned}$$

By the factorisation theorem $(\hat\Sigma,\hat\mu)$ is sufficient.

For X and X' two observations, let's note $\hat u=\hat\Sigma+(\hat\mu-\mu)(\hat\mu-\mu)^T$, $\hat u'=\hat\Sigma'+(\hat\mu'-\mu)(\hat\mu'-\mu)^T$

$$\frac{f_\theta^n(X)}{f_\theta^n(X')}=\exp\{-\frac{n}{2}\mathrm{Tr}\Sigma^{-1}\left(\hat u-\hat u'\right)\}$$

Let's suppose that this quantity doesn't depend on θ. so

$$\mathrm{Tr}\Sigma_1^{-1}\left(\hat u-\hat u'\right)=\mathrm{Tr}2\Sigma_1^{-1}\left(\hat u-\hat u'\right)=0$$

But $\hat u-\hat u'$ is symmetric non negative, so there exist P a invertible matrix and $D=\mathrm{diag}(a_1,...a_n)$ such that it equals PDP^{-1}.

Let $\Sigma=P\mathrm{diag}(b_1,...,b_n)P^{-1}$ where $b_i=\frac{1}{a_n}$ if $a_n\neq 0$, 1 otherwise.

then $\mathrm{Tr}(\Sigma^{-1}PDP^{-1})=\sum a_i^2=0$, which means $\hat u=\hat u'$. eg:

$$(\forall\mu)\;\hat\Sigma+(\hat\mu-\mu)(\hat\mu-\mu)^T=\hat\Sigma'+(\hat\mu'-\mu)(\hat\mu'-\mu)^T$$

Letting $\mu=\hat\mu$, we have $\hat\Sigma=\hat\Sigma'+(\hat\mu'-\hat\mu)(\hat\mu'-\hat\mu)^T$

Letting $\mu=\hat\mu'$, we have $\hat\Sigma=\hat\Sigma'-(\hat\mu'-\hat\mu)(\hat\mu'-\hat\mu)^T$

We conclude that $\hat\Sigma=\hat\Sigma'$ and $||\hat\mu-\hat\mu'||^2=\mathrm{Tr}(\hat\mu'-\hat\mu)(\hat\mu'-\hat\mu)^T=0$, eg $\hat\mu=\hat\mu'$. And therfore $(\hat\Sigma,\hat\mu)$ is minimal.

Question 6

The log-likelihood function:

$$\begin{aligned}
\mathcal{L}(\theta;x) &= \log(\Pi_i f(x_i|\theta)) && \text{because iid} && (10)\\
&= \log\frac{1}{(\sigma\sqrt{2\pi})^n}e^{-\sum_i\frac{(x_i-\mu)^2}{2\sigma^2}} && && (11)\\
&= -n\log(\sqrt{2\pi})-n\log\sigma-\sum_i\frac{(x_i-\mu)^2}{2\sigma^2} && && (12)\\
& && && (13)
\end{aligned}$$

We maximize first in μ.

$$\mu = \arg\max -\sum_i (x_i - \mu)^2$$

Since the function is concave in μ, we find the optimum by setting the first derivative to 0, ie $\mu = \bar{x}$
We now maximize in σ by setting the first derivative to 0 and verifying thatthe second derivative is negative

$$\frac{d\mathcal{L}}{d\sigma}(\mu = \bar{x}, \sigma) = -\frac{n}{\sigma} + \frac{\sum_i (x_i - \bar{x})^2}{\sigma^3} = 0 \Rightarrow \sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$$
$$\frac{d^2\mathcal{L}}{d\sigma^2}(\mu = \bar{x}, \sigma^2 = S_n^2) = \frac{n}{\sigma^2} - 3\frac{\sum_i (x_i - \bar{x})^2}{\sigma^4} = \frac{n}{S_n^2} - \frac{3n}{S_n^2} < 0$$

MLE

$$\theta = (\bar{x}, \frac{1}{n}\sum_i (x_i - \bar{x})^2) = (\bar{x}, S_n^2)$$

Question 7

Let's denote

$$\theta := (p_l, \mu_l, \Sigma_l)_l$$

Estimation

$$\begin{aligned}
&\mathcal{L}^n(X, L; \theta) && (14)\\
&= \prod_i \mathcal{L}(X_i, L_i; \theta) && (15)\\
&= \prod_i \sum_l \mathbb{P}(L_i = l) f(X_i, \mu_l, \Sigma_l) 1_{\{L=l\}} && (16)\\
&= \prod_i \sum_l {}_l 1_{\{L=l\}} \frac{e^{-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j)}}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma_j}} && (17)\\
&= \prod_i \exp \sum_l 1_{L=l} (-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) - \frac{n \log(2\pi)}{2} - \frac{\log \det \Sigma_j}{2} + \log p_l) && (18)\\
&= \exp \sum_{i,l} 1_{L=l} (-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) - \frac{n \log(2\pi)}{2} - \frac{\log \det \Sigma_j}{2} + \log p_l) && (19)
\end{aligned}$$

$$\begin{aligned}
Q(\theta, \theta') &= \mathbb{E}^{\theta'}[\log \mathcal{L}^n(X, L|\theta)|X] && (20)\\
&= \sum_{i,l} \mathbb{E}^{\theta'}[1_{L_i=l} (-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) - \frac{n \log(2\pi)}{2} - \frac{\log \det \Sigma_j}{2} + \log p_l | X] && (21)\\
&= \sum_{i,l} \mathbb{P}^{\theta'}(L_i = l|X) (-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) - \frac{n \log(2\pi)}{2} - \frac{\log \det \Sigma_j}{2} + \log p_l) && (22)\\
&= \sum_{i,l} T_{i,l}^{\theta'} (-\frac{1}{2}(X_i - \mu_j)^T \Sigma_j^{-1} (X_i - \mu_j) - \frac{n \log(2\pi)}{2} - \frac{\log \det \Sigma_j}{2} + \log p_l) && (23)\\
& && (24)
\end{aligned}$$

Where:

$$T_{i,l}^{\theta'} := \mathbb{P}^{\theta'}(L_i = l|X_i) \tag{25}$$
$$= \frac{f(L_i = l, X = X_i|\theta')}{f(X = X_i; \theta')} \tag{26}$$
$$= \frac{\mathbb{P}(L = l|\theta')f(X = X_i|L = l; \theta')}{\sum_k \mathbb{P}(L = k|\theta')f(X = X_i|L = k; \theta')} \tag{27}$$

Maximization

We can optimize first in p, μ and then Σ

1.

$$p^* := \arg\max_{p, \sum_l p_l = 1} Q(\theta, \theta') \tag{28}$$
$$= \arg\max_p \sum_{i,l} T_{i,l}^{\theta'} \log p_l \tag{29}$$

Using lagrange multiplier:

$$p^* := \arg\max_{p, \lambda \geq 0} \sum_{i,l} T_{i,l}^{\theta'} \log p_l - \lambda(1 - \sum_l p_l) \tag{30}$$
$$= g(p, \lambda) \tag{31}$$

For $l = 1..n$:

$$0 = \frac{\partial g}{\partial p_l} = \frac{\sum_i T_{i,l}^{\theta'}}{p_l} + \lambda \Rightarrow p_l = -\frac{\sum_i T_{i,l}^{\theta'}}{\lambda}$$

and since $\sum_l p_l = 1$, $\lambda = -\sum_{i,l} T_{i,l}^{\theta'} = -n$ and therefore:

$$p_l^* = \frac{1}{n} \sum_i T_{i,l}^{\theta'}$$

2. The optimization in μ and Σ looks like the optimization in question 6.

$$\mu_l^*, \Sigma_l^* = \arg\max_{\mu_l, \Sigma_l} Q(\theta, \theta') \tag{32}$$
$$= \arg\max_{\mu, \Sigma} \sum_i T_{i,l}^{\theta'} (-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu) - \frac{\log|\det \Sigma|}{2}) \tag{33}$$

If we write $\Sigma = \text{diag}(\sigma_1, ...\sigma_n) > 0$ and $\mu = (\mu_k)_k$ we have:

$$\mu_l^*, \Sigma_l^* = \arg\max_{\mu, \Sigma} \sum_i T_{i,l}^{\theta'} \left(-\frac{1}{2} \sum_k \frac{1}{\sigma_k}((X_i)_k - \mu_k)^2 - \frac{1}{2} \sum_k log|\sigma_k| \right) \tag{34}$$
$$= \arg\min_{\mu, \Sigma} \sum_{k,i} T_{i,l}^{\theta'} \frac{1}{\sigma_k}((X_i)_k - \mu_k)^2 + \sum_k (\sum_i T_{i,l}^{\theta'}) log|\sigma_k| \tag{35}$$

The program is quadratic and concave in μ, so by setting the first derivative to 0:

$$\mu_l^* = \frac{\sum_i T_{i,l}^{\theta'} X_i}{\sum_i T_{i,l}^{\theta'}}$$

3. We can optimize on each σ_k independenly by linearity, σ_k^* is the solution to:

$$\frac{\partial}{\partial \sigma_k}\sum_i T_{i,l}^{\theta'}\frac{1}{\sigma_k}((X_i)_k-\mu_k^*)^2-(\sum_i T_{i,l}^{\theta'})log(\sigma_k) = -(\sum_i T_{i,l}^{\theta'}((X_i)_k-\mu_k^*)^2)\frac{1}{\sigma_k^2}+(\sum_i T_{i,l}^{\theta'})\frac{1}{\sigma_k}=0$$

so:

$$\sigma_k^* = \frac{\sum_i T_{i,l}^{\theta'}((X_i)_k-\mu_k^*)^2}{\sum_i T_{i,l}^{\theta'}}$$

(we can calculate the second derivative to prove that this is indeed a minimum like we did in question 6).

Question 8

-

$$\begin{aligned}
\mathbb{P}(\hat{\theta}\le x) &= \mathbb{P}\{\max_i x_i \le x\} && (36)\\
&= \mathbb{P}(\cap_i\{x_i\le x\}) && (37)\\
&= \Pi_i \mathbb{P}(x_i\le x) \quad \text{By independence} && (38)\\
&= \min{(1, \left(\frac{x}{\theta}\right)^n)} && (39)\\
&= \int_{\mathbb{R}} n\frac{y^{n-1}}{\theta^n}1_{0\le y\le\theta}1_{y\le x}\mathrm{dy} && (40)\\
&= \int^x f(y)\mathrm{dy} \quad \text{where } f(y)=n\frac{y^{n-1}}{\theta^n}1_{0\le y\le \theta} && (41)
\end{aligned}$$

  f is the density of $\hat{\theta}$ w.r.t to lebesgue measure.

-

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \int_0^\theta yn\frac{y^{n-1}}{\theta^n}\mathrm{dy} && (42)\\
&= \frac{n}{n+1}\theta && (43)
\end{aligned}$$

  So $\hat{\theta}$ is biased for $\theta\neq 0$.

Question 9

1. By definition of conditional probability:

$$\mathcal{L}(\theta;x)=\mathcal{L}(\theta;x_1|x_2,...)\mathcal{L}(\theta;x_2|x_3,...)...\mathcal{L}(\theta;x_n)$$

$$\mathbb{E}\log\mathcal{L}(x;\theta)=\sum_i\mathbb{E}\log\mathcal{L}(x_i;\theta|x_{i+1}...x_n)=-\sum_i H(x_i|x_{i+1}...x_n)$$

   If the x_i are iid:

$$\mathbb{E}\log\mathcal{L}(x;\theta)=-\sum_i H(x_i)$$

2.

$$\begin{aligned}
H(X) - H(X|Y) &= \mathbb{E}log(f(Y)/f(X,Y)) && (44)\\
&\leq \log \mathbb{E}\frac{f(Y)}{f(X,Y)} && \text{By concavity of log} \quad (45)\\
&= \log \int \frac{f(Y)}{f(X,Y)} f(X,Y) && (46)\\
&= \log 1 = 0 && (47)
\end{aligned}$$

Question 10

$$g(\beta) = \sum (y_i - x_i^T\beta)^2$$

$$f(\beta) = \sum (y_i - x_i^T\beta)^2) + \lambda||\beta||^2 + g(\beta) = \lambda||\beta||^2$$

$$\begin{aligned}
\nabla_\beta f &= \sum_i -2(y_i - x_i^T\beta)x_i + 2\lambda\beta && (48)\\
&= 2(\lambda\beta - \sum_i (y_i - x_i^T\beta)x_i) && (49)\\
&= 2((\lambda I_n + \sum_i x_i x_i^T)\beta + \sum_i y_i x_i) && (50)
\end{aligned}$$

The hessian of f is $F := 2(\lambda I_n + \sum_i x_i x_i^T)$. F is symetric and its eigen values are those of $\sum_i x_i x_i^T$ offset by λ. For λ large enough ($\lambda > ||\sum_i x_i x_i^T||_\infty$), the eigen values of F are all positive, and therefore f is strictly convexe and admit at most one global minimum.

In addition, there is a solution iff $\nabla f = 0$ has a solution, and the solution happens to be the minimum. Which is the case for

$$\beta = \frac{1}{2}F^{-1}\sum y_i x_i = (\lambda I_n + \sum_i x_i x_i^T)^{-1} \sum_i y_i x_i$$

Question 11

Let's consider that $\mathcal{L}(\beta) \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda})$, and $\mathcal{L}(Y|\beta) = \mathcal{N}(X^T\beta, \Sigma^2)$ where $\Sigma^2 = \mathrm{diag}(\sigma_i^2)_i$

then

$$f(\beta) = cte\, e^{-\frac{\lambda}{2\sigma^2}\beta^T\beta}$$

$$f(Y|\beta) = cte\, e^{-\frac{1}{2\sigma^2}\sum_i (y_i - x_i^T\beta)^2}$$

$$\sum (y_i - x_i^T\beta)^2 + \lambda||\beta||^2 = cte - \log(e^{-\sum_i \frac{1}{\sigma^2}(y_i - x_i^T\beta)^2} e^{-\frac{\lambda}{\sigma^2}\beta^T\beta}) = cte - log(f(Y|\beta)g(\beta))$$

Minimizing for ridge regression is the same as maximizing the posterior distribution of β: $\arg\max_\beta h(\beta|Y) = \arg\max_\beta f(Y|\beta)g(\beta) = \arg\max_\beta \sum (y_i - x_i^T\beta)^2 + \lambda||\beta||^2$

Question 12

Let $\hat{X} = (X^l)_{l \in \mathbb{N}^p : |l| \leq k}$

$$Y = poly(X) + \epsilon = \beta \hat{X} + \epsilon$$

$\beta = (\sum_i \hat{x}_i \hat{x}_i^T)^{-1} \sum_i y_i \hat{x}_i$ by using the precedent questions.