Exercise Sheet #1

November 14, 2018

Remark 1. *The below computations become a little easier (but esentially the same), if the logarithm is applied to the maximized quantity first such that the exponential disappears and the product is turned into a sum.*

1 Exercise 1a)

We have to show that the maximum-likelihood (ML) estimator for w is

$$w_{\mathrm{ML}} := \operatorname{argmax}_{w \in \mathbb{R}^F} p(y|X, w) = (\phi_x \phi_x^\intercal)^{-1} \phi_x y. \tag{1}$$

Proof. The likelihood is given by

$$\begin{aligned}
\mathcal{L}(w) &:= p(y|X, w) = \mathcal{N}(y; \phi_x^\intercal w, \sigma^2 I_n) \\
&= \frac{1}{\sqrt{(2\pi)^n \, |\sigma^2 I_n|}} \exp\left(-\frac{1}{2}(y - \phi_x^\intercal w)^\intercal \sigma^{-2} I_n (y - \phi_x^\intercal w)\right) \qquad (2) \\
&= C \exp\left(-\frac{1}{2}(y - \phi_x^\intercal w)^\intercal \sigma^{-2} I_n (y - \phi_x^\intercal w)\right) \\
&= C \exp\left(-\frac{1}{2\sigma^2}(y - \phi_x^\intercal w)^\intercal (y - \phi_x^\intercal w)\right) \\
&= C \exp\left(-\frac{1}{2\sigma^2}(y - g(w))^\intercal (y - g(w))\right), \qquad (3)
\end{aligned}$$

with

$$\begin{aligned}
g : \mathbb{R}^F &\to \mathbb{R}^n, && (4) \\
w &\mapsto \phi_x^\intercal w, && (5)
\end{aligned}$$

where $C > 0$ is some constant independent of w. Now, define

$$\begin{aligned}
h : \mathbb{R}^n &\to \mathbb{R} && (6) \\
v &\mapsto \exp\left(-\frac{1}{2\sigma^2}(y - v)^\intercal (y - v)\right), && (7)
\end{aligned}$$

such that

$$\mathcal{L}(w) = (h \circ g)(w), \quad \forall w \in \mathbb{R}^F. \tag{8}$$

By the chain rule for gradients (a special case of the chain rule for total derivatives),

$$\nabla \mathcal{L}(w) = J_g^\intercal(w) \nabla h(g(w)) \tag{9}$$

where $J_g \in \mathbb{R}^{n \times F}$ denotes the Jacobian matrix of g, i.e.

$$J_g = \begin{pmatrix} \frac{\partial g_1}{\partial w_1} & \cdots & \frac{\partial g_1}{\partial w_F} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial w_1} & \cdots & \frac{\partial g_n}{\partial w_F} \end{pmatrix} \overset{(5)}{=} \phi_x^\intercal. \tag{10}$$

Since

$$h(v) \overset{(7)}{=} \exp\left(l(v)\right) \tag{11}$$

with

$$\begin{aligned} l : \mathbb{R}^F &\rightarrow \mathbb{R}, && (12) \\ v &\mapsto -\frac{1}{2\sigma^2}(y^\intercal y - 2y^\intercal v + vv^\intercal). && (13) \end{aligned}$$

Since

$$\nabla l(v) = -\frac{1}{2\sigma^2}\left(2y - 2v\right) = -\frac{1}{\sigma^2}(y - v), \tag{14}$$

the chain rule for gradients implies that

$$\nabla h(v) = \exp\left(l(v)\right) \nabla l(v) = -\frac{1}{\sigma^2} \exp\left(l(v)\right)(y - v). \tag{15}$$

Insertion of (15) and (10) into (9) yields

$$\begin{aligned} \nabla \mathcal{L}(w) &= \phi_x \left[-\frac{1}{\sigma^2} \exp(l(v))(y - g(w)) \right] && (16) \\ &= C\phi_x(y - \phi_x^\intercal w) && (17) \\ &= C(\phi_x y - \phi_x^\intercal w), && (18) \end{aligned}$$

which is, due to $C > 0$, equal to 0 if and only if $w = \left(\phi_x \phi_x^\intercal\right)^{-1} \phi_x y$. As $\sigma^{-2} I_n$ is as a precision matrix positive definite, it can be seen from (2) that it is a minimum and not a maximum.

□

2 Exercise 1b)

We have to show that the maximum a posteriori (MAP) estimator of the posterior

$$p(w|y,\phi_x) = \mathcal{N}\left(w; \left(\Sigma^{-1} + \sigma^{-2}\phi_x^\intercal\phi_x\right)^{-1}\left(\Sigma^{-1}\mu + \sigma^{-2}\phi_x y\right), \left(\Sigma^{-1} + \sigma^{-2}\phi_X^\intercal\phi_X\right)^{-1}\right) \tag{19}$$

is identical to the posterior mean, i.e.

$$w_{MAP} = \left(\Sigma^{-1} + \sigma^{-2}\phi_x^\intercal\phi_x\right)^{-1}\left(\Sigma^{-1}\mu + \sigma^{-2}\phi_x y\right). \tag{20}$$

We show the following stronger statement (from which the above statement follows by inserting $m := \left(\Sigma^{-1} + \sigma^{-2}\phi_x^\intercal\phi_x\right)^{-1}\left(\Sigma^{-1}\mu + \sigma^{-2}\phi_x y\right)$ and $V := \left(\Sigma^{-1} + \sigma^{-2}\phi_X^\intercal\phi_X\right)^{-1}$ into the below theorem).

Theorem 1. *Let $F \in \mathbb{N}$, $m \in \mathbb{R}^F$ and $V \in \mathbb{R}^{F\times F}$. Then,*

$$m = \operatorname{argmax}_{z\in\mathbb{R}} \mathcal{N}(z; m, V). \tag{21}$$

Proof (Theorem 1.) Recall that

$$\begin{aligned}
p(z) &:= \mathcal{N}(z; m, V) && (22)\\
&= \frac{1}{\sqrt{(2\pi)^n|V}} \exp\left(-\frac{1}{2}(z-m)^\intercal V^{-1}(z-m)\right) && (23)\\
&= C\exp\left(-\frac{1}{2}(z-m)^\intercal V^{-1}(z-m)\right) && (24)\\
&= C\exp\left(-\frac{1}{2}g(z)\right), && (25)
\end{aligned}$$

with constant $C > 0$ independent of z and

$$\begin{aligned}
g : \mathbb{R}^F &\to \mathbb{R}, && (26)\\
z &\mapsto (z-m)^\intercal V^{-1}(z-m). && (27)
\end{aligned}$$

By the chain rule for gradients, we have

$$\nabla p(z) = -\frac{C}{2}\exp\left(-\frac{1}{2}g(z)\right)\nabla g(z). \tag{28}$$

To compute ∇g, let

$$\begin{aligned}
h : \mathbb{R}^F \times \mathbb{R}^F &\to \mathbb{R}, && (29)\\
(v, z) &\mapsto v^T z, && (30)
\end{aligned}$$

such that $g(z) = h((z-m), V^{-1}(z-m))$ and

$$\frac{\partial h(v,z)}{dv} = z \tag{31}$$

$$\frac{\partial h(v,y)}{dz} = v \tag{32}$$

Now, by application of the multivariate chain rule to h, we deduce

$$\nabla g(z) = \underbrace{\frac{\partial h}{\partial v}((z-m), V^{-1}(z-m))}_{=V^{-1}(z-m)} + V^{-\intercal} \underbrace{\frac{\partial h}{\partial z}((z-m), V^{-1}(z-m))}_{=(z-m)} \tag{33}$$

$$= \left(V^{-1} + V^{-T}\right)(z-m) = 2V^{-1}(z-m). \tag{34}$$

Hence, by (28) and $-\frac{C}{2}\exp\left(-\frac{1}{2}g(z)\right) > 0$,

$$\nabla p(z) = 0 \iff \nabla g(z) = 0 \iff (z-m) = 0 \iff z = m. \tag{35}$$

Hence, z is the unique extremal point of m. As V^{-1} is as a precision matrix positive definite, g and p thereby monotonously increases in every dimension of z. Hence m is a minimum of $\mathcal{N}(z; m, V)$. $\square$