1. Given

$$f(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

where $x, \mu \in \mathbb{R}^k$, Σ is a k-by-k positive definite matrix and $|\Sigma|$ is its determinant. Show that $\int_{\mathbb{R}^k} f(x)\, dx = 1$.

Let $y = x - u \Rightarrow dy = dx$

Then $= \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2} y^T \Sigma^{-1} y} dy$

$\because$ Σ is a k-by-k positive definite matrix

$\therefore$ By Cholesky decomposition, we can get $\Sigma = LL^T$, where L is an invertible lower triangular matrix.

Then $|\Sigma| = |LL^T| = |L||L^T| = |L|^2 \Rightarrow |L| = |\Sigma|^{\frac{1}{2}}$

Let $y = L^{-1}(x-u) \Rightarrow x = u + Ly$, $dx = |L|\, dy$.

And
$$\begin{aligned}(x-u)^T\Sigma^{-1}(x-u) &= (x-u)^T(LL^T)^{-1}(x-u)\\ &= (x-u)^T(L^{-T}L^{-1})(x-u)\\ &= (L^{-1}(x-u))^T(L^{-1}(x-u))\\ &= y^Ty = \|y\|^2\end{aligned}$$

Hence,
$$\begin{aligned}\int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-u)^T\Sigma^{-1}(x-u)} dx &= \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}\|y\|^2} |L|\, dy\\ &= \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \cdot |\Sigma|^{\frac{1}{2}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}\|y\|^2} dy\\ &= \frac{1}{\sqrt{(2\pi)^k}} \prod_{i=1}^{k} \left(\underline{\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy}\right)\end{aligned}$$

by Gauss integral, $\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$

$$= \frac{1}{\sqrt{(2\pi)^k}} (\sqrt{2\pi})^k = 1$$

2. Let A, B be n-by-n matrices and x be a n-by-1 vector.

(a) Show that $\frac{\partial}{\partial A}\mathrm{trace}(AB) = B^T$.

(b) Show that $x^TAx = \mathrm{trace}(xx^TA)$.

(b) Derive the maximum likelihood estimators for a multivariate Gaussian.

(a) Suppose $(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$

$\mathrm{trace}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}B_{ki}$

$\frac{\partial}{\partial A_{pq}}\mathrm{trace}(AB) = \frac{\partial}{\partial A_{pq}}\left(\sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}B_{ki}\right)$, $\because \frac{\partial A_{ik}}{\partial A_{pq}} = \begin{cases} 1 \ ; \text{if } i=p \text{ and } k=q \\ 0 \ ; \text{o.w.} \end{cases}$

$1 \le p, q \le n$

$= 0 \cdot \sum_{i\ne p}^{n}\sum_{k\ne q}^{n} B_{ki} + 1 \cdot B_{qp}$

$= B_{qp}$

Therefore $\frac{\partial}{\partial A}\mathrm{trace}(AB) = B^T$. *

(b) $\because x^TAx$ is a scalar, $\therefore x^TAx = \mathrm{trace}(x^TAx)$

Suppose A, B, C are $n \times n$ matrix.

Note that $\mathrm{trace}(ABC) = \mathrm{trace}(BCA) = \mathrm{trace}(CAB)$

Hence, $\mathrm{trace}(x^TAx) = \mathrm{trace}(xx^TA)$. *

(c) Suppose $X_i \overset{iid}{\sim} N_n(\mu, \Sigma)$ for $i = 1, 2, \dots, m$, where $\mu \in \mathbb{R}^n$ and

Σ is a $n\times n$ matrix.

$f(x_i; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$

$L(\mu, \Sigma) = \prod_{i=1}^{m} f(x_i; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{m}{2}}}\prod_{i=1}^{n}\exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$

$\ell(\mu, \Sigma) = \ln L(\mu, \Sigma) = -\frac{mn}{2}\ln(2\pi) - \frac{m}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$

$\frac{\partial}{\partial \mu}\ell(\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^{m}\frac{\partial}{\partial\mu}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$

$= -\frac{1}{2}\sum_{i=1}^{m}\left(-2\Sigma^{-1}(x_i-\mu)\right)$

$= \Sigma^{-1}\sum_{i=1}^{m}(x_i-\mu)$

Let $\frac{\partial}{\partial\mu}\ell(\mu,\Sigma) = 0 \Rightarrow \sum_{i=1}^{m}(x_i-\mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{m} x_i}{m}$ *

$\because \sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$ is a scalar, $\therefore$ by (b), $\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = \mathrm{trace}\left(\sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$

Hence, $\ell(\mu,\Sigma) = -\frac{mn}{2}\ln(2\pi) - \frac{m}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$

$= -\frac{mn}{2}\ln(2\pi) - \frac{m}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{trace}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$

Let $S = \sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$

$$\frac{\partial}{\partial \Sigma^{-1}} l(u, \Sigma) = \frac{m}{2}\Sigma - \frac{1}{2}S$$

Let $\frac{\partial}{\partial \Sigma^{-1}} l(u, \Sigma) = 0 \Rightarrow \frac{m}{2}\Sigma - \frac{1}{2}S = 0 \Rightarrow \hat{\Sigma} = \frac{S}{m} = \frac{1}{m}\sum_{i=1}^{n}(x_i - u)(x_i - u)^T$ #

#3.

1. softmax 的對稱性使不同參數組合給出相同輸出，這會如何影響可辨識性與正則化設計？
2. cross-entropy 對於標註噪聲（label noise）敏感度如何？什麼情況需要改用其他損失或加權？