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 $\int_{\mathbb{R}^k} f(x)\, dx = 1$.

1° Since Σ is a positive definite matrix, and is defined on $M_{\mathbb{R}^k \times \mathbb{R}^k}$, by using Cholesky deconposition, let $\Sigma = LL^T$ — ① which L is lower triangle matrix

2° Consider exponent in exponential by using ① :

$$-\frac{1}{2}(x-\mu)^T (LL^T)^{-1} L^{-1}(x-\mu)$$

$$= -\frac{1}{2}(x-\mu)^T (L^{-1})^T L^{-1}(x-\mu)$$

$$= -\frac{1}{2}(L^{-1}\cdot(x-\mu))^T (L^{-1}(x-\mu)), \text{ let } y = L^{-1}\cdot(x-\mu)$$

$$= -\frac{1}{2}\cdot(y\cdot y)$$

$$= -\frac{1}{2}\|y\|^2, \quad y \in \mathbb{R}^k \times \mathbb{R}^1$$

3° Jacobian and integral transform

$y = L^{-1}(x-\mu)$, $x = \mu + Ly$

then $\frac{dx}{dy} = \det(L) = |\Sigma|^{\frac{1}{2}}$

4° For integrating:

$z \sim N(\mu \in \mathbb{R}^k, \Sigma \in M_{\mathbb{R}_k \times \mathbb{R}_k})$

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

$$\int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k \cdot |\Sigma|}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) dx$$

$$= \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k \cdot |\Sigma|}} \cdot \exp\left(-\frac{1}{2}\|y\|^2\right) \cdot |\Sigma|^{\frac{1}{2}} \cdot dy$$

$$= \int_{\mathbb{R}^k} \frac{\cancel{|\Sigma|}^{\frac{1}{2}}}{\sqrt{(2\pi)^k \cdot \cancel{|\Sigma|}}} \cdot \exp\left(-\frac{1}{2}\left\|\sum_{i=1}^{k} y_i\right\|^2\right) dy_i, \ y_i \in \mathbb{R}^1$$

$$= \frac{1}{\sqrt{(2\pi)^k}} \cdot \prod_{i=1}^{k} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}y_i^2\right) \cdot dy_i$$

$$= \frac{1}{\sqrt{(2\pi)^k}} \cdot \left(\prod_{i=1}^{k} \sqrt{2\pi}\right) = 1 \quad \#$$

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

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

(b) Show that $x^T A x = \text{trace}(x x^T A)$.

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

(a)

Claim $\frac{\partial tr(UV)}{\partial X} = \frac{\partial tr(U_c V)}{\partial X} + \frac{\partial tr(U V_c)}{\partial X}$

which U_c, V_c are constant matrix not relation to X

1° $tr(UV) = \sum_k (UV)_{kk} = \sum_k \sum_l u_{kl} v_{lk}$

2° $\frac{\partial tr(UV)}{\partial X} = \sum_k \sum_l \frac{\partial u_{kl} v_{lk}}{\partial x_{ij}}$

$= \sum_k \sum_l \left(u_{kl} \frac{\partial v_{lk}}{\partial x_{ij}} + v_{lk} \frac{\partial u_{kl}}{\partial x_{ij}} \right)$

$= \sum_k \sum_l u_{kl} \frac{\partial v_{lk}}{\partial x_{ij}} + \sum_k \sum_l v_{lk} \frac{\partial u_{kl}}{\partial x_{ij}}$

find that front u_{kl} and bak v_{lk} not relative to x_{ij}

$= \left(\frac{\partial tr(U_c V)}{\partial X} + \frac{\partial tr(U V_c)}{\partial X} \right)_{ij}$

then we claimed it

for $\frac{\partial tr(AB)}{\partial A}$:

$$\frac{\partial tr(AB)}{\partial A} = \frac{\partial tr(A_c B)}{\partial A} + \frac{\partial (A B_c)}{\partial A}$$

$$= \sum_{k=1}^{n} \sum_{l=1}^{n} a_{kl} \frac{\partial b_{kl}}{\partial a_{ij}} + \sum_{k=1}^{n} \sum_{l=1}^{n} b_{kl} \frac{\partial a_{kl}}{\partial a_{ij}}$$

which $\frac{\partial b_{kl}}{\partial a_{ij}} = 0$

$$= \sum_{k=1}^{n} \sum_{l=1}^{n} b_{kl} \frac{\partial a_{kl}}{\partial a_{ij}}$$

use kronecker delta

$$\delta_{ij} = \begin{cases} 1 & , \text{if } i=j \\ 0 & , \text{if } i \neq j \end{cases}$$

$$= \sum_{k=1}^{n} \sum_{l=1}^{n} b_{kl} \delta_{il} \delta_{jk}$$

$$= b_{ji}$$

$$= B^T$$ #1

参考:https://ccjou.wordpress.com/2013/06/03/%E8%B7%A1%E6%95%B8%E8%88%87%E8%A1%8C%E5%88%97%E5%BC%8F%E7%9A%84%E5%B0%8E%E6%95%B8/
tr-5,tr-6

(b)

Extend $x^T A x$ =

1° let e_j be canonical bias vectors :

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},\quad \cdots,\quad e_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

then $x = \sum_{j=1}^{n} x_j e_j$

2° $$x^T A x = \left(\sum_{j=1}^{n} x_j e_j^T\right) A \left(\sum_{j=1}^{n} x_j e_j\right)$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k e_j^T A e_k$$

Extend $tr(Axx^T)$

1° $$xx^T = \left(\sum_{j=1}^{n} x_j e_j\right)\left(\sum_{k=1}^{n} x_k e_k^T\right)$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k e_j e_k^T$$

2° $$tr(Axx^T) = tr\left(A \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k e_j e_k^T\right)$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} x_j x_k \, tr(A e_j e_k^T)$$

Notice that:

$$tr(Ae_je_k^T) = tr(e_k^T A e_j), \quad tr(ABC) = tr(CAB)$$

$$= e_k^T A e_j, \quad \dim M_{(1\times n)} \times M_{(n\times n)} \times M_{(n\times 1)} = M_{(1\times 1)} = \text{scalar}$$

$$= e_j^T A e_k$$

thus: $x^T A x = tr(Axx^T) = \sum_{j,k} x_j x_k e_j^T A e_k$

and use the fact:

$$tr(AB) = tr(BA)$$

$$\Rightarrow tr(Axx^T) = tr(xx^T A)$$

then get $x^T A x = tr(xx^T A)$ #

參考 https://math.stackexchange.com/questions/766049/a-proof-of-xtax-mathrmtraxxt

(c)

1° let $X_1, X_2, \cdots X_n \sim N(\mu, \Sigma)$, Σ is positive detrzite

and its likelihood $L(\mu, \Sigma) = \prod_{i=1}^{n} f(x_i \mid \mu, \Sigma)$, $\Sigma \in M_{k\times k}$

2°
$$L(\mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$
$$= \frac{1}{(2\pi)^{\frac{nk}{2}} \cdot |\Sigma|^{\frac{n}{2}}} \cdot \prod_{i=1}^{n} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$
$\sum_{i=1}^{n}$

3° let $l(\mu, \Sigma) = \ln L(\mu, \Sigma)$
$$= -\frac{nk}{2} \ln(2\pi) - \frac{n}{2} \ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1} (x_i-\mu)$$

Notice that
$$\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu) = tr\left(\Sigma^{-1} \sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\right)$$
by 2 (b) $x^T A x = tr(A x x^T) = tr(x x^T A)$

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

then get: $l(\mu, \Sigma) = -\frac{nk}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2} tr(\Sigma^{-1} S(\mu))$

4° $\frac{\partial}{\partial \mu} l(\mu, \Sigma) = 0$:

$$\frac{\partial}{\partial \mu} l(\mu, \Sigma) = -\frac{1}{2} \frac{\partial}{\partial \mu} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu), \text{ for } \forall i$$

remark for 2.(a), let $(x_i - \mu) = \vec{x}$, $A = \Sigma^{-1}$

$$\Rightarrow \frac{\partial}{\partial \mu} tr(A x b^T) = A^T b$$

$$\frac{\partial}{\partial \mu} tr(A(xb^T + bx^T)) = 2A^T b, \quad (\Sigma^{-1})^T = \Sigma^{-1}$$

$$\text{Thus } \frac{\partial}{\partial \mu} l(\mu, \Sigma) = -\frac{1}{2} \cdot 2(\Sigma^{-1})^T (x_i - \mu), \forall i$$

$$= -\Sigma^{-1}(x_i - \mu), \forall i = 0$$

$$\Rightarrow \Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad \#$$

5° $\frac{\partial}{\partial \Sigma} l(\mu, \Sigma) = 0$

$$\frac{\partial}{\partial \Sigma} l(\mu, \Sigma) = \underbrace{\frac{\partial}{\partial \Sigma}\left(-\frac{n}{2} \ln|\Sigma|\right)}_{①} + \underbrace{\frac{\partial}{\partial \Sigma}\left(-\frac{1}{2} tr(\Sigma^{-1} S(\mu))\right)}_{②}$$

① : use the fact which in some books

$$\frac{\partial}{\partial A} \ln|A| = (A^{-1})^T$$

$$\Rightarrow \frac{\partial}{\partial \Sigma}\left(-\frac{n}{2}\ln|\Sigma|\right) = -\frac{n}{2}\Sigma^{-1}$$

②:

by 2(a).

$$\frac{\partial}{\partial A} tr(A^{-1}B) = -(A^{-1}BA^{-1})^T$$

$$\Rightarrow \frac{\partial}{\partial \Sigma}\left(-\frac{1}{2} tr(\Sigma^{-1}S(\mu))\right) = -\frac{1}{2}\cdot\left(-(\Sigma^{-1}S(\mu)\Sigma^{-1})\right)$$

① + ②

$$= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}S(\mu)\Sigma^{-1}$$

$$= 0$$

$$\Rightarrow \Sigma^{-1}S(\mu)\Sigma^{-1} = n\Sigma^{-1}$$

$$\Sigma^{-1}S(\mu) = n$$

$$S(\mu) = n\Sigma$$

$$\text{So } \hat{\Sigma}_{MLE} = \frac{1}{n}S = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T$$

#1

6° for log-likelihood $(\hat{\mu}, \hat{\Sigma})$, check that second derivative (Hessian) is negative:

① for μ:

$$\frac{\partial l}{\partial \mu} = \Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu) = n \Sigma^{-1} (\bar{x} - \mu)$$

$$\frac{\partial^2 l}{\partial \mu^2} = -n \Sigma^{-1} < 0 \text{, ok}$$

↖ positive definite

② for Σ:

$$\frac{\partial l}{\partial \Sigma} = -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} S \Sigma^{-1}$$

$$\hat{\Sigma}_{MSE} = \frac{1}{n} S$$

Use Hessian matrix to check:

$$l''(H) = -\frac{n}{2} tr(\Sigma^{-1} H \Sigma^{-1} H)$$

$$= -\frac{n}{2} \| \Sigma^{-\frac{1}{2}} H \Sigma^{-\frac{1}{2}} \|^2 \leq 0 \text{, ok}$$

so $(\hat{\mu}, \hat{\Sigma})$ is MLE of Multi... Gaussia...

3,

什麼情況下比較適合用 LDA，而什麼情況又要用 GDA ？

當各類的資料分布差異比較大、邊界呈現非線性時就應該改用 GDA？

另外，如果樣本數太少導致協方差矩陣估不準，這兩種方法會不會失效或需要做正則化處理呢？