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 $y = x - \mu$

$$\Rightarrow I = \int_{\mathbb{R}^k} f(x)\, dx = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2} y^T \Sigma^{-1} y}\, dy$$

$\because \Sigma$ is positive definite

$\therefore \exists$ P is orthogonal matrix, Λ is diagonal matrix s.t $\Sigma = P\Lambda P^T$

$$\Rightarrow \Sigma^{-1} = (P\Lambda P^T)^{-1} = (P^T)^{-1}\Lambda^{-1}P^{-1} = P\Lambda^{-1}P^T \quad (\because P^T = P^{-1})$$

$$\Rightarrow y^T\Sigma^{-1}y = y^T(P\Lambda^{-1}P^T)y$$

Let $z = P^T y$ and knew that $|\det(P^T)| = 1$ (i.e, $dz = dy$)

$$\Rightarrow y^T\Sigma^{-1}y = (Pz)^T(P\Lambda^{-1}P^T)(Pz) = z^T P^T P \Lambda^{-1} P^T P z$$

$$= z^T \Lambda^{-1} z$$

$$\Rightarrow z^T\Lambda^{-1}z = \sum_{i=1}^{k} \frac{z_i^2}{\lambda_i} \quad \text{where } \lambda_i \text{ is } \Lambda\text{'s eigenvalues} \quad i = 1, \cdots, k$$

Since $|\Sigma| = |P\Lambda P^T| = |P||\Lambda||P^T| = 1 \cdot |\Lambda| \cdot 1 = \prod_{i=1}^{k} \lambda_i$,

$$I = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k \prod_{i=1}^{k}\lambda_i}} e^{-\frac{1}{2}\sum_{i=1}^{k}\frac{z_i^2}{\lambda_i}}\, dz$$

$$= \frac{1}{\prod_{i=1}^{k}\sqrt{2\pi\lambda_i}} \int_{\mathbb{R}^k} \left(\prod_{i=1}^{k} e^{-\frac{z_i^2}{2\lambda_i}}\right) dz$$

$$= \left(\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\lambda_i}}\right)\left(\prod_{i=1}^{k} \int_{-\infty}^{\infty} e^{-\frac{z_i^2}{2\lambda_i}}\, dz_i\right)$$

Note $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$

Then, $\int_{-\infty}^{\infty} e^{-\frac{z_i^2}{2\lambda_i}} dz_i = \sqrt{\frac{\pi}{\left(\frac{1}{2\lambda_i}\right)}}$ (Let $a = \frac{1}{2\lambda_i}$)

$$= \sqrt{2\pi\lambda_i}$$

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

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 $x^T A x = \text{trace}(x x^T A)$.
   (b) Derive the maximum likelihood estimators for a multivariate Gaussian.

(a) Note that $\text{trace}(AB) = \sum_{i=1}^{n} (AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ji}$

Suppose we partial a_{kl}

$$\left(\frac{\partial}{\partial A} \text{trace}(AB)\right)_{kl} = \frac{\partial}{\partial a_{kl}} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ji} \right)$$

$$= \frac{\partial}{\partial a_{kl}} (a_{kl} b_{lk})$$

$$= b_{lk}$$

Thus, $\frac{\partial}{\partial A} \text{trace}(AB) = B^T$ #

(b) Since $x^T A x$ is scalar,

$x^T A x = \text{trace}(x^T A x)$

Note that $\text{trace}(AB) = \text{trace}(BA)$

$$\Rightarrow \begin{cases} tr(x^T(Ax)) = tr((Ax)x^T) \\ Tr((xx^T)A) = tr(A(xx^T)) \end{cases}$$

Thus, $x^T A x = tr(x x^T A)$ #

(c) Note $L(\mu,\Sigma;X) = \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))$

where $X=\{x_1,x_2,\cdots,x_N\}$

Let $\ell = \ln L$

$$\ell(\mu,\Sigma) = \ln\left[((2\pi)^k|\Sigma|)^{-\frac{N}{2}} \exp(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))\right)$$

$$= -\frac{Nk}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

Then, $\nabla_\mu \ell = -\frac{1}{2}\sum_{i=1}^{N} \nabla_\mu((x_i-\mu)^T\Sigma^{-1}(x_i-\mu))$

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

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

Let $\nabla_\mu \ell = 0$

$\because \Sigma$ is positive definite $\therefore \exists\, \Sigma^{-1}$ s.t $\sum_{i=1}^{N}(x_i-\mu) = 0$

$\Rightarrow \sum_{i=1}^{N} x_i - N\mu = 0$

$\Rightarrow N\mu = \sum_{i=1}^{N} x_i$

Then, $\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$\ln(\mu,\Sigma) = -\frac{Nk}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N} tr((x_i-\mu)(x_i-\mu)^T\Sigma^{-1})$ by (b)

$$= -\frac{Nk}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2} tr((\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T)\Sigma^{-1})$$

Let $S_\mu = \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$ and $C = -\frac{Nk}{2}\ln(2\pi)$

$\ln(\mu,\Sigma) = C - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}tr(S_\mu\Sigma^{-1})$

Note that $\frac{\partial}{\partial A}\ln|A| = (A^{-1})^T$, $\frac{\partial}{\partial A} tr(BA) = B^T$

Let $U = \Sigma^{-1}$

$$l = C + \frac{N}{2} \ln|U| - \frac{1}{2} tr(S_\mu U)$$

$$\Rightarrow \frac{\partial}{\partial U} = \frac{N}{2}(U^{-1})^T - \frac{1}{2}(S_\mu)^T$$

$\because$ U and S_μ are symmetric $\quad \therefore (U^{-1})^T = \Sigma^T = \Sigma$, $(S_\mu)^T = S_\mu$

$$\Rightarrow \frac{\partial}{\partial U} = \frac{N}{2}\Sigma - \frac{1}{2} S_\mu$$

Let $\frac{\partial}{\partial U} = 0 \quad \Rightarrow \frac{N}{2}\Sigma - \frac{1}{2} S_\mu = 0$

$$\Rightarrow N\Sigma = S_\mu$$

Thus, $\hat{\Sigma}_{MLE} = \frac{S_\mu}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})(x_i - \hat{\mu}_{MLE})^T \#$$

3、 Questions

Compared to Softmax regression, which natively handles multi-class problems, in what situations does One-vs-Rest perform poorly?
Does it have any advantages that Softmax does not?