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

Pf: $\because$ Σ is positive definite.

$\therefore$ By Cholesky decompositive, $\exists !$ 對角線都嚴格大於0的下三角矩陣 L s.t.

$$\Sigma = LL^T. \qquad (L^{-1} \text{ exists})$$

Let $y = L^{-1}(x-\mu)$, then $x = \mu + Ly$, and $dx = |\det L|\,dy = |\Sigma|^{\frac{1}{2}}dy$

$\Rightarrow -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}(Ly)^T(LL^T)^{-1}(Ly) = -\frac{1}{2}y^TL^T(L^T)^{-1}L^{-1}Ly = -\frac{1}{2}y^Ty = -\frac{1}{2}\|y$

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

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

Consider $\int_{\mathbb{R}^k} e^{-\|z\|^2}dz$

By Fubini's Theorem, $\int_{\mathbb{R}^k} e^{-\|z\|^2}dz = \pi^{\frac{k}{2}}$

Let $z = \frac{1}{\sqrt{2}}y$, then $\|z\| = \frac{1}{\sqrt{2}}\|y\| \Rightarrow \|z\|^2 = \frac{1}{2}\|y\|^2$ and $dz = \left(\frac{1}{\sqrt{2}}\right)^k dy$

$$\therefore \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k}} e^{-\frac{1}{2}\|y\|^2}dy = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k}} e^{-\|z\|^2}(\sqrt{2})^k dz$$

$$= \frac{1}{\sqrt{\pi^k}} \int_{\mathbb{R}^k} e^{-\|z\|^2}dz$$

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

$$= 1$$

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

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

(a) Def : $\left[\frac{\partial}{\partial A} C\right]_{ij} = \frac{\partial}{\partial A_{ij}} C$, i.e., $\frac{\partial}{\partial A} C = \begin{bmatrix} \frac{\partial C}{\partial A_{11}} & \frac{\partial C}{\partial A_{12}} & \cdots & \frac{\partial C}{\partial A_{1n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial C}{\partial A_{n1}} & \frac{\partial C}{\partial A_{n2}} & \cdots & \frac{\partial C}{\partial A_{nn}} \end{bmatrix}$

$$\mathrm{trace}(AB) = \sum_{p=1}^{n} (AB)_{pp} = \sum_{p=1}^{n}\left(\sum_{q=1}^{n} A_{pq} B_{qp}\right)$$

$$\left[\frac{\partial}{\partial A}\mathrm{trace}(AB)\right]_{ij} = \frac{\partial}{\partial A_{ij}}\mathrm{trace}(AB) = \frac{\partial}{\partial A_{ij}}\left[\sum_{p=1}^{n}\sum_{q=1}^{n} A_{pq}B_{qp}\right]$$

$$= \sum_{p=1}^{n}\sum_{q=1}^{n}\frac{\partial}{\partial A_{ij}} A_{pq}B_{qp}$$

$$= \sum_{p=1}^{n}\sum_{q=1}^{n}\delta_{ip}\delta_{jq}B_{qp} = B_{ji}$$

$$\Rightarrow \frac{\partial}{\partial A}\mathrm{trace}(AB) = \begin{bmatrix} B_{11} & B_{21} & \cdots & B_{n1} \\ B_{12} & B_{22} & \cdots & B_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ B_{1n} & B_{2n} & \cdots & B_{nn} \end{bmatrix} = B^T$$

(b)

$$x^T A x = [x_1 \ x_2 \cdots x_n]\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= x_1\sum_{p=1}^{n} x_p A_{p1} + x_2\sum_{p=1}^{n} x_p A_{p2} + \cdots + x_n\sum_{p=1}^{n} x_p A_{pn}$$

$$= \sum_{q=1}^{n}\left(x_q\sum_{p=1}^{n} x_p A_{pq}\right) = \sum_{q=1}^{n}\sum_{p=1}^{n} x_p x_q A_{pq}$$

$$x x^T = [x_1 \ x_2 \cdots x_n]\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1^2 & x_1x_2 & \cdots & x_1x_n \\ x_1x_2 & x_2^2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_1x_n & x_2x_n & \cdots & x_n^2 \end{bmatrix} =: X$$

$$\mathrm{trace}(xx^TA) = \sum_{q=1}^{n} (\Sigma A)_{qq} = \sum_{q=1}^{n} \sum_{p=1}^{n} \Sigma_{qp} A_{pq}$$

$$= \sum_{q=1}^{n} \sum_{p=1}^{n} x_q x_p A_{pq} \quad \because \Sigma_{ij} = x_i x_j$$

$$= \sum_{q=1}^{n} \sum_{p=1}^{n} x_p x_q A_{pq}$$

$$\therefore x^TAx = \mathrm{trace}(xx^TA)$$

(c) Let $x^{(1)}, x^{(2)}, \cdots, x^{(N)} \in \mathbb{R}^n$ and $x^{(k)} \sim N(\mu, \Sigma)$

Likelihood function $L(\mu, \Sigma) = \prod_{i=1}^{N} P(x^{(i)})$

$$= \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)}$$

We want to find μ, Σ s.t $L(\mu, \Sigma)$ is max.

Define $\ell(\mu, \Sigma) := \ln L(\mu, \Sigma)$

$$= \ln\left\{ \left[\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}}\right]^N e^{-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)} \cdot \cdots \cdot e^{-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)} \right\}$$

$$= -N \ln\left[(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}\right] + \sum_{i=1}^{N} -\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1} (x^{(i)}-\mu)$$

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

$$\because \frac{\partial}{\partial \mu}\left[(x-\mu)^T \Sigma^{-1} (x-\mu)\right] = -2\Sigma^{-1}(x-\mu)$$

$$\therefore \frac{\partial}{\partial \mu} \ell(\mu, \Sigma) = -\frac{1}{2} \sum_{i=1}^{N} \left[-2\Sigma^{-1}(x^{(i)}-\mu)\right]$$

Let $\frac{\partial}{\partial \mu} \ell(\mu, \Sigma) = 0$, i.e., $\sum_{i=1}^{N} \Sigma^{-1}(x^{(i)}-\mu) = 0$

$$\Rightarrow \mu N = \sum_{i=1}^{N} x^{(i)} \Rightarrow \mu = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

$$\because \frac{\partial}{\partial \Sigma} \ln|\Sigma| = \Sigma^{-1} \quad \text{and} \quad \frac{\partial}{\partial \Sigma} (x-\mu)^T \Sigma^{-1} (x-\mu) = \frac{\partial}{\partial \Sigma} \left[\operatorname{trace}\left((x-\mu)(x-\mu)^T \Sigma^{-1} \right) \right]$$

$$= \frac{\partial}{\partial \Sigma} \left[\operatorname{trace}\left(\Sigma^{-1} (x-\mu)(x-\mu)^T \right) \right]$$

$$= -\Sigma^{-1} (x-\mu)(x-\mu)^T \Sigma^{-1}$$

$$\therefore \frac{\partial}{\partial \Sigma} l(\mu, \Sigma) = -\frac{N}{2} \Sigma^{-1} - \frac{1}{2} \left\{ \sum_{i=1}^{N} \left[-\Sigma^{-1} (x^{(i)}-\mu)(x^{(i)}-\mu)^T \Sigma^{-1} \right] \right\}$$

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

Let $\frac{\partial}{\partial \Sigma} l(\mu, \Sigma) = 0$, then $\sum_{i=1}^{N} \Sigma^{-1} (x^{(i)}-\mu)(x^{(i)}-\mu)^T = N$

$$\Rightarrow \Sigma = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

3. Unanswered Questions

There are unanswered questions from the lecture, and there are likely more questions we haven't covered.

- Take a moment to think about these questions.
- Write down the ones you find important, confusing, or interesting.
- You do **not** need to answer them—just state them clearly.

The exponentially family 的形式是 $P(y;\eta) = b(y)\exp(\eta^T y - a(y))$

(可用來說明在一些 distribution 下，hypothesis function 為什麼是如此假設)

Q：為什麼形式是這樣？

從哪裡推導來的？

背後還有什麼更深的理論支持嗎？