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

where $x, \mu \in \mathbb{R}^k$, Σ is a k by k positive define matrix and $|\Sigma|$ is its determinant.

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

Since Σ is a positive-definite matrix.

There exist a orthogonal matrix $P \in M_{k\times k}(\mathbb{R})$ and diagonal matrix $D \in M_{k\times k}(\mathbb{R})$

s.t $\Sigma = PDP^* = PDP^T = PD^{\frac{1}{2}}D^{\frac{1}{2}}P^T = AA^T$, where $A = PD^{\frac{1}{2}}$, $D^{\frac{1}{2}}_{ij} = \sqrt{D_{ij}}$

$|\Sigma| = |AA^T|$, since A is invertible, $|\Sigma| = |A||A^T| = |A|^2$

let $y = A^{-1}(x-\mu) \Rightarrow x = \mu + Ay$, $dx = |A|\,dy$

and $(x-\mu)^T \Sigma^{-1}(x-\mu) = (Ay)^T (AA^T)^{-1}(Ay) = y^T A^T (A^T)^{-1} A^{-1} Ay = y^T y$

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

$$= \frac{1}{\sqrt{(2\pi)^k |A|^2}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}(y_1^2 + y_2^2 + \cdots + y_k^2)} |A|\,dy_1\,dy_2 \cdots dy_k$$

$$= \frac{1}{\sqrt{(2\pi)^k |A|^2}} |A| \prod_{i=1}^{k} \int_{\mathbb{R}} e^{-\frac{1}{2}y_i^2}\,dy_i$$

$$= \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{\frac{-y_i^2}{2}}\,dy_i$$

$$= 1$$

2.

Let A, B be $n \times n$ matrices and x be $n \times 1$ vector.

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

pf :

$$\frac{\partial}{\partial A} \operatorname{trace}(AB) = \frac{\partial \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ji}}{\partial \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}} = \sum_{i=1}^{m} \sum_{j=1}^{n} b_{ji} = B^T$$

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

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

$$= \begin{bmatrix} \langle x, A^{(1)} \rangle & \cdots & \langle x, A^{(n)} \rangle \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= x_1 \langle x, A^{(1)} \rangle + \cdots + x_n \langle x, A^{(n)} \rangle$$

$$= x_1 (a_{11} x_1 + a_{21} x_2 + \cdots + a_{n1} x_n) + \cdots + x_n (a_{1n} x_1 + a_{2n} x_2 + \cdots + a_{nn} x_n)$$

$$= \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} a_{ji} x_j \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j a_{ji}$$

$$\operatorname{trace}(x x^T A) = \operatorname{trace}\left(\begin{bmatrix} x_1 x_1 & x_1 x_2 & \cdots & x_1 x_n \\ x_2 x_1 & & & \\ \vdots & & & \\ x_n x_1 & x_n x_2 & \cdots & x_n x_n \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i x_j)(a_{ji})$$

Hence, $x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j a_{ji} = \operatorname{trace}(x x^T A)$

(c)

Derive the maximum likelihood estimator for multivariate Gaussian.

Let $x_1, \cdots, x_n \overset{iid}{\sim} f(x)$

$$L(x_1, \cdots, x_n \mid \mu, \Sigma) = \prod_{i=1}^{n} (2\pi)^{\frac{-k}{2}} |\Sigma|^{\frac{-1}{2}} e^{\frac{-1}{2}(x_i - \mu)^T \Sigma (x_i - \mu)}$$

$$\log L = \ell(\mu, \Sigma) = -\frac{nk}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

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

$$= \frac{-1}{2} \sum_{i=1}^{n} -(\Sigma^{-1})^T x_i - (\Sigma^{-1}) x_i + \frac{\partial}{\partial \mu} (\operatorname{trace}(\mu \mu^T \Sigma^{-1}))$$

$$= \frac{-1}{2} \sum_{i=1}^{n} -2 \Sigma^{-1} x_i + 2\mu \Sigma^{-1}$$

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

$$= \Sigma^{-1} \left(\sum_{i=1}^{n} x_i - n\mu \right) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

$$= \frac{-n}{2} (\Sigma^{-1})^T - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \Sigma} \operatorname{trace}\left(\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T \right)$$

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

$$= \frac{-n}{2} (\Sigma^{-1})^T + \frac{1}{2} \sum_{i=1}^{n} (\Sigma^{-1})^T (x_i - \mu)^T (x_i - \mu) (\Sigma^{-1})^T = 0$$

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

$$\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^T (x_i - \hat{\mu}_{MLE}) = \hat{\Sigma}_{MLE}$$

3. Unanswer Questions

上課有提到 exponential family 的 canonical form 求 $E(T(Y))$

2.1 Densities and Parameters

Let μ be a measure on $\mathbb{R}^n$, let $h : \mathbb{R}^n \to \mathbb{R}$ be a nonnegative function, and let $T_1, \ldots, T_s$ be measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$. For $\eta \in \mathbb{R}^s$, define

$$A(\eta) = \log \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x). \tag{2.1}$$

Whenever $A(\eta) < \infty$, the function p_η given by

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x), \qquad x \in \mathbb{R}^n, \tag{2.2}$$

integrates to one; that is, $\int p_\eta\, d\mu = 1$. So, this construction gives a family of probability densities indexed by η. The set

$$\Xi = \{\eta : A(\eta) < \infty\}$$

is called the *natural parameter space*, and the family of densities $\{p_\eta : \eta \in \Xi\}$ is called an *s-parameter exponential family* in *canonical form*.

2.2 Differential Identities

In canonical exponential families it is possible to relate moments and cumulants for the statistics $T_1, \ldots, T_s$ to derivatives of A. The following theorem plays a central role.

Theorem 2.4. *Let Ξ_f be the set of values for $\eta \in \mathbb{R}^s$ where*

$$\int |f(x)| \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x) < \infty.$$

Then the function

$$g(\eta) = \int f(x) \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x)$$

is continuous and has continuous partial derivatives of all orders for $\eta \in \Xi_f^\circ$ (the interior of Ξ_f). Furthermore, these derivatives can be computed by differentiation under the integral sign.

A proof of this result is given in Brown (1986), a monograph on exponential families with statistical applications. Although the proof is omitted here, key ideas from it are of independent interest and are presented in the next section. As an application of this result, if $f = 1$, then $\Xi_f = \Xi$, and, by (2.1),

$$g(\eta) = e^{A(\eta)} = \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x).$$

Differentiating this expression with respect to η_j, which can be done under the integral if $\eta \in \Xi^\circ$, gives

$$\begin{aligned} e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} &= \int \frac{\partial}{\partial \eta_j} \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x) \\ &= \int T_j(x) \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x)\, d\mu(x). \end{aligned}$$

Using the definition (2.2) of p_η, division by $e^{A(\eta)}$ gives

$$\frac{\partial A(\eta)}{\partial \eta_j} = \int T_j(x) p_\eta(x)\, d\mu(x).$$

This shows that if data X has density p_η with respect to μ, then

$$E_\eta T_j(X) = \frac{\partial A(\eta)}{\partial \eta_j} \tag{2.4}$$

for any $\eta \in \Xi^\circ$.

Derivatives at the origin give cumulants, and thus cumulants for the sum T will equal the sum of the corresponding cumulants of $Y_1, \ldots, Y_n$. This is a well-known result for the mean and variance.

If X has density from a canonical exponential family (2.2), and if $T = T(X)$, then T has moment generating function

$$\begin{aligned} E_\eta e^{u \cdot T(X)} &= \int e^{u \cdot T(x)} e^{\eta \cdot T(x) - A(\eta)} h(x)\, d\mu(x) \\ &= e^{A(u+\eta) - A(\eta)} \int e^{(u+\eta) \cdot T(x) - A(u+\eta)} h(x)\, d\mu(x), \end{aligned}$$

provided $u + \eta \in \Xi$. The final integrand is $p_{u+\eta}$, which integrates to one. So, the moment generating function is $e^{A(u+\eta) - A(\eta)}$, and the cumulant generating function is

$$K_T(u) = A(u + \eta) - A(\eta).$$

Taking derivatives, the cumulants for T are

$$\kappa_{r_1, \ldots, r_s} = \frac{\partial^{r_1}}{\partial \eta_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial \eta_s^{r_1}} A(\eta).$$

Example 2.10. If X has the Poisson distribution with mean λ, then

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x \log \lambda - \lambda}, \qquad x = 0, 1, \ldots.$$

The mass functions for X form an exponential family, but the family is not in canonical form. The canonical parameter here is $\eta = \log \lambda$. The mass function expressed using η is

$$P(X = x) = \frac{1}{x!} \exp[\eta x - e^\eta], \qquad x = 0, 1, \ldots,$$

and so $A(\eta) = e^\eta$. Taking derivatives, all of the cumulants of $T = X$ are $e^\eta = \lambda$.

Example 2.11. The class of normal densities formed by varying μ with σ^2 fixed can be written as

$$p_\mu(x) = \exp\left[\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}.$$

These densities form an exponential family with $T(x) = x$, canonical parameter $\eta = \mu/\sigma^2$, and $A(\eta) = \sigma^2\eta^2/2$. The first two cumulants are $\kappa_1 = A'(\eta) = \sigma^2\eta = \mu$ and $\kappa_2 = A''(\eta) = \sigma^2$. Because A is quadratic, all higher-order cumulants, $\kappa_3, \kappa_4, \ldots$, are zero.