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

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

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

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

Let $I = \int_{\mathbb{R}^k} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} dx$, we need to prove $I = \sqrt{(2\pi)^k|\Sigma|}$

Let $y = x-\mu$, $\det(J_1) = \det(I) = 1$ Hence $dx = dy$, $I = \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^T\Sigma^{-1}y} dy$

$\because \Sigma$ is symmetrical positive determined matrix $\therefore \Sigma^{-1}$ is also symmetrical positive determined matrix;

then $\exists$ a orthogonal matrix P ($P^TP = I$) and a diag. matrix D s.t. $\Sigma^{-1} = PDP^T$. $D = \mathrm{diag}(\lambda_1, \lambda_2, \cdots \lambda_k)$

Let $z = P^Ty$, then $y = Pz$. $\det(J_2) = \det(P) = \pm 1$. Hence $dy = dz$

$y^T\Sigma^{-1}y = (Pz)^T(PDP^T)(Pz) = z^TP^TPDP^TPz = z^TIDIz = z^TDz = \sum_{i=1}^{k} \lambda_i z_i^2$

$I = \int_{\mathbb{R}^k} e^{-\frac{1}{2}\sum_{i=1}^{k}\lambda_i z_i^2} dz = \int_{\mathbb{R}^k} \prod_{i=1}^{k} e^{-\frac{1}{2}\lambda_i z_i^2} dz_1 \dots dz_k = \prod_{i=1}^{k} \int_{-\infty}^{\infty} e^{-\frac{\lambda_i}{2} z_i^2} dz_i$

use Gaussian Integral: $\int_{-\infty}^{\infty} e^{-au^2} du = \sqrt{\frac{\pi}{a}}$, $a = \frac{\lambda_i}{2}$,

$\Rightarrow I = \prod_{i=1}^{k} \sqrt{\frac{\pi}{\frac{\lambda_i}{2}}} = \prod_{i=1}^{k} \sqrt{\frac{2\pi}{\lambda_i}} = \sqrt{\frac{(2\pi)^k}{\prod_{i=1}^{k}\lambda_i}}$, $\prod_{i=1}^{k} \lambda_i = \det(D) = \det(P^T\Sigma^{-1}P) = \det(P^T)\det(\Sigma^{-1})\det(P) = \det(\Sigma^{-1})$

$\because \det(\Sigma^{-1}) = \frac{1}{\det(\Sigma)} = \frac{1}{|\Sigma|}$, then $I = \sqrt{\frac{(2\pi)^k}{\frac{1}{|\Sigma|}}} = \sqrt{(2\pi)^k|\Sigma|}$

$\int_{\mathbb{R}^k} f(x)\, dx = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \cdot I = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \cdot \sqrt{(2\pi)^k|\Sigma|} = 1$ #

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

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

(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) Let elements in A, B are A_{ij} and B_{ij}

$(AB)_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$, $\text{trace}(AB) = \sum_{i=1}^{n}(AB)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}$

$\frac{\partial}{\partial A_{kl}}\text{trace}(AB) = \frac{\partial}{\partial A_{kl}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}\right) \Rightarrow \frac{\partial}{\partial A_{kl}}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ji}\right) = B_{lk}$

By defn. of the matrix derivative, $\frac{\partial f(A)}{\partial A}$ is a matrix whose (k,l)-th element is $\frac{\partial f(A)}{\partial A_{kl}}$.

Therefore, the (k,l)-th element of $\frac{\partial}{\partial A}\text{trace}(AB)$ is B_{lk}

A matrix C whose (k,l)-element is equal to the (l,k)-th element of another matrix B is precisely the defn of the transpose, i.e. $C = B^T$. $\left(\frac{\partial}{\partial A}\text{trace}(AB)\right)_{kl} = B_{lk} = (B^T)_{kl}$

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

(b) $x : n\times 1$, $A : n\times n$, $x^T A x : (1\times n)(n\times n)(n\times 1) \rightarrow 1\times 1$

$x x^T A : (n\times 1)(1\times n)(n\times n) \rightarrow n\times n$

$tr(CDE) = tr(DEC) = tr(ECD)$ (By cyclic Property of Trace)

Let $C = x$, $D = x^T$, $E = A$

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

For $\forall$ scalar s, $tr(s) = s$

$\therefore tr(x^T A x) = x^T A x$

$\Rightarrow tr(x x^T A) = tr(x^T A x) = x^T A x$ #

(c) PDF : $p(x_i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$

$L(\mu, \Sigma; D) = \prod_{i=1}^{N} p(x_i \mid \mu, \Sigma)$. In order to facilitate the derivation, we maximize its log. func., that is, the log. similarity func. $l(\mu,\Sigma) = \log L(\mu, \Sigma; D)$.

Then $l(\mu,\Sigma) = \log\left(\prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\right) = \sum_{i=1}^{N} \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}\right)$

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

Hence $l(\mu,\Sigma) = -\frac{Nn}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$.

We find the μ that maximizes l by taking the gradient of μ and setting it to zero. $\nabla_\mu l(\mu,\Sigma) = 0$

We need to compute $\nabla_\mu\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$ use $\nabla_v(v^T C v) = 2Cv$ (when C is sysmetric) & $\nabla_v(b^T C v) = C^T b$.

Then we get $\nabla_\mu((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)) = -2\Sigma^{-1}(x_i-\mu)$ subst. into gradient. $\Rightarrow \nabla_\mu l(\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^{N}(-2\Sigma^{-1}(x_i-\mu)) = \sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu) = \Sigma^{-1}\left(\sum_{i=1}^{N}x_i - N\mu\right)$.

Let the gradient $\Sigma^{-1}\left(\sum_{i=1}^{N}x_i - N\mu\right) = 0 \Rightarrow \Sigma\Sigma^{-1}\left(\sum_{i=1}^{N}x_i - N\mu\right) = \Sigma 0 \Rightarrow \sum_{i=1}^{N}x_i - N\mu = 0 \Rightarrow N\mu = \sum_{i=1}^{N}x_i$

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

We rewrite the log-likelihood func. in terms of S, and use $\log|\Sigma| = \log|S^{-1}| = -\log|S|$

$\ell(\mu, S) = C + \frac{N}{2}\log|S| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T S(x_i-\mu)$ use (b) $v^T M v = tr(vv^T M)$, then $\sum_{i=1}^{N}(x_i-\mu)^T S(x_i-\mu) = \sum_{i=1}^{N} tr((x_i-\mu)(x_i-\mu)^T S)$

use $\sum tr(A_i) = tr(\sum A_i)$: $\sum tr(A_i) = tr([\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T]S)$.

let $S_\mu = \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$, ℓ becomes to $\ell(\mu,S) = C + \frac{N}{2}\log|S| - \frac{1}{2}tr(S_\mu S)$

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

$\because S = \Sigma^{-1}$ & S_μ are sysemtric matrices $\therefore (S^{-1})^T = S^{-1} = \Sigma$ & $(S_\mu)^T = S_\mu$

$\Rightarrow \frac{N}{2}\Sigma - \frac{1}{2}S_\mu = 0 \quad \Rightarrow \Sigma = \frac{1}{N}S_\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$

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

$\therefore \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}x_i$ and $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$ #

3、Question: In notes mention that "Gaussion Simple Bayesian" assume a diagonal covariance matrix, Does it mean that we can find all features $(x_1, x_2, \cdots, x_d)$ are independent of each other in special cases? I'd like to know what types of examples would perform better or worse under this assumption.