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.>

1° For $\Sigma_{k\times k}$ is symmetric and positive definite, there exists an invertible matrix A s.t. $\Sigma = AA^T$, moreover $|\Sigma| = (\det A)^2$

Now, we define the new variable $y = A^{-1}(x-\mu) \Rightarrow x = \mu + Ay$

Moreover, we have the Jacobian determinant of this transformation is:

$$dx = |\det A|\,dy = \sqrt{|\Sigma|}\,dy$$

2° Substituting $x=\mu+Ay$ into our exponent:

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = y^TA^T(AA^T)^{-1}Ay = y^Ty = \|y\|^2.$$

3° $$\int_{\mathbb{R}^k} f(x)\,dx = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\int_{\mathbb{R}^k}\exp\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx$$

$$\overset{\boxed{\text{By } 2°}}{=} \frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\int_{\mathbb{R}^k}\exp\left(\frac{-1}{2}\|y\|^2\right)\boxed{|\det A|}\,dy$$

$$= \frac{|\det A|}{\sqrt{(2\pi)^k|\Sigma|}}\int_{\mathbb{R}^k}\exp\left(\frac{-1}{2}\underbrace{\|y\|^2}_{(y_1^2+y_2^2+\cdots+y_k^2)}\right)dy$$

※ $\exp\left(\frac{-1}{2}\|y\|^2\right) = \prod_{j=1}^{k}\exp\left(\frac{-1}{2}y_j^2\right)$

where, $$\int_{\mathbb{R}^k}\exp\left(\frac{-1}{2}\|y\|^2\right)dy = \boxed{\prod_{j=1}^{k}}\int_{-\infty}^{\infty}e^{\frac{-1}{2}y_j^2}\,dy_j = (\sqrt{2\pi})^k = (2\pi)^{\frac{k}{2}}$$

Therefore, $$\int_{\mathbb{R}^k} f(x)\,dx = \frac{\cancel{|\Sigma|}}{\cancel{\sqrt{(2\pi)^k}}\cancel{|\Sigma|}}\cancel{(2\pi)^{\frac{k}{2}}} = 1 \quad \blacksquare$$

$$\boxed{\begin{array}{l} ※ \int_{-\infty}^{\infty}e^{-ax^2}dx = \sqrt{\frac{\pi}{a}} \\ \text{Then, } \int_{-\infty}^{\infty}e^{\frac{-1}{2}y_j^2}dy_j = \sqrt{2\pi}\end{array}}$$

By Caculus 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 $x^TAx = \text{trace}(xx^TA)$.

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

(a)

<pf.> 1° Suppose that $A=[a_{ij}]$; $B=[b_{ij}]$, then $(AB)_{ii} = \sum_{k=1}^{n} a_{ik}b_{ki}$. Therefore, $tr(AB) = \sum_{i=1}^{n}\sum_{k=1}^{n}a_{ik}b_{ki}$

2° We take the partial derivative for a_{mn}.

We can see that the only term which depend on a_{mn} is when $i=m$; $k=n$.

Thus, the derivative: $\frac{\partial}{\partial a_{mn}}tr(AB) = b_{mn}$

3° The gradiant $\frac{\partial}{\partial A}tr(AB)$ is the matrix with (m,n)-entry is exactly this derivative:

i.e. $\left[\frac{\partial}{\partial A}tr(AB)\right]_{mn} = \frac{\partial}{\partial a_{mn}}tr(AB) = b_{nm}$

However, the matrix of B^T has entries $(B^T)_{mn} = b_{nm}$.

Therefore, we prove that $\frac{\partial}{\partial A}tr(AB) = B^T$ ▨

<method 2> "Differential + Frobenius inner product"

1° Let $f(x) = tr(AB)$, where B is a constant

Then, the differential of f is: $df = d\,tr(AB) = tr(dAB) = tr(dA\,B)$

(for B doesn't depent on A)

2° Using the cyclic property of the trace: Frobenius inner product.

Using $tr(dA\,B) = tr(B\,dA) = tr((B^T)^T dA) = \langle B^T, dA\rangle_F$, where $\langle X, Y\rangle_F = tr(X^TY)$

3° By the definition of the matrix gradient: $df = \langle \nabla_A f, dA\rangle_F$

However, we also know that $df = \langle B^T, dA\rangle_F$

Since this must hold for all variation dA, we conclude: $\nabla_A f = \frac{\partial}{\partial A}tr(AB) = B^T$ ▨

(b) 1° X^TAX is a scaler (ie X^TAX is a 1×1 matrix)

$\forall S$ is scaler, we have $S = tr(S)$

For the trace of a 1×1 matrix is just the value itself (i.e. $X^TAX = tr(X^TAX)$)

2° Using the cyclic property of trace:

(for example: $tr(PQR) = tr(QRP) = tr(RPQ)$), Here $P = X^T_{1\times n}$; $Q = A_{n\times n}$; $R = X_{n\times 1}$

Then, we have $tr(X^TAX) = tr(AXX^T) = tr(XX^TA)$

3° The matrix XX^T is an $n\times n$ outer product, so the final expression $tr(XX^TA)$ is exactly the same with the original scalar X^TAX

Therefore, $X^TAX = tr(XX^TA)$ ▨

(c) 1° Suppose that we have m independent samples:

$\{X_1, X_2, \ldots\ldots, X_m\} \in \mathbb{R}^k$, where $x_i \sim \mathcal{N}(\mu, \Sigma)$, with unknown mean vector $\mu \in \mathbb{R}^k$ and the co-variance matrix $\Sigma \in \mathbb{R}^{k\times k}$ (symmetric positive define).

2° The probability density function for one sample:

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

Moreover, the likelihood function for the full data-set: $\mathcal{L}(\mu,\Sigma) = \prod_{i=1}^{m} p(x_i|\mu,\Sigma)$

$$\rightarrow l(\mu,\Sigma) = \frac{-m}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

☆ Define the scatter matrix $S(\mu) = \sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T$

Using the result of (b), then we have $\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = tr(\Sigma^{-1}S(\mu))$

Therefore, $\ell(\mu,\Sigma) = \frac{-m}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}S(\mu))$

① 3° $\frac{\partial \ell}{\partial \mu} = \frac{-1}{2}\left(\frac{\partial}{\partial \mu}\mathrm{tr}(\Sigma^{-1}S(\mu))\right) = \Sigma^{-1}\left(\sum_{i=1}^{m} x_i - m\mu\right)$ ← Expanding $S(\mu)$ and the differentiating.

Suppose that $\Sigma^{-1}\left(\sum_{i=1}^{m} x_i - m\mu\right) = 0$, then $\sum_{i=1}^{m} x_i - m\mu = 0$, we have $\boxed{\hat{\mu} = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i}$ ∎

4° ② $\frac{\partial}{\partial \Sigma}\ln|\Sigma| = \Sigma^{-1}$; $\frac{\partial}{\partial \Sigma}\mathrm{tr}(\Sigma^{-1}S) = -\Sigma^{-T}S\Sigma^{-T}$

then, $\frac{\partial \ell}{\partial \Sigma} = \frac{-m}{2}\Sigma^{-T} + \frac{1}{2}\Sigma^{-T}S(\hat{\mu})\Sigma^{-T}$

Suppose that $\frac{-m}{2}\Sigma^{-T} + \frac{1}{2}\Sigma^{-T}S(\hat{\mu})\Sigma^{-T} = 0$, moreover, multiply on the left and right by Σ^{T}

we can get $S(\hat{\mu}) = m\Sigma$

Thus, $\boxed{\hat{\Sigma} = \frac{1}{m}S(\hat{\mu}) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^T}$ ∎

3. Unanswered Questions

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

Q: If there is a multiclass classification, how do we find the boundary?