

## ML HW5

Date: 10/01/2025

Week: 5

Author: Alvin B. Lin

Student ID: 112652040

### Problem 1: Integral of Multivariable Normal Distribution

Given

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

where  $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^k$ ,  $\boldsymbol{\Sigma}$  is a  $k$ -by- $k$  positive definite matrix and  $|\boldsymbol{\Sigma}|$  is its determinant. Show that  $\int_{\mathbb{R}^k} f(\mathbf{x}) d\mathbf{x} = 1$ .

**Solution:**

We first do a substitution:

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}; \quad \Omega = \mathbb{R}^k \implies \mathbb{R}^k = \Omega'.$$

We need the following lemma:

#### Theorem 1: Symmetric $\iff$ Orthogonally Diagonalisable

Given a matrix  $\mathbf{A} \in \mathbb{M}_{k \times k}$ , then it follows that

$$\mathbf{A} \text{ is symmetric } \iff \mathbf{A} = \mathbf{Q}\mathbf{U}\mathbf{Q}^T$$

for some **orthogonal**  $\mathbf{Q} \in \mathbb{M}_{k \times k}$ , **diagonal**  $\mathbf{D} \in \mathbb{M}_{k \times k}$ .

Combine **Theorem 1** and the **positive definite** property on  $\boldsymbol{\Sigma}$ , we know that:

$$\boldsymbol{\Sigma} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T = (\mathbf{Q}\sqrt{\mathbf{D}})(\mathbf{Q}\sqrt{\mathbf{D}})^T := (\boldsymbol{\Sigma}^{1/2})(\boldsymbol{\Sigma}^{1/2})^T,$$

where  $\mathbf{Q}$  is orthogonal,  $\mathbf{D}$  is a diagonal matrix with all entries positive,  $\sqrt{\mathbf{D}}^2 = \mathbf{D}$ .

So now we have our exponent part becomes:

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}(\boldsymbol{\Sigma}^{-1/2}\mathbf{y})^T (\boldsymbol{\Sigma}^{-1/2}\mathbf{y}) = -\frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{y}\|^2$$

By letting  $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$ , we get  $d\mathbf{z} = |\boldsymbol{\Sigma}^{-1/2}|d\mathbf{y} = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}d\mathbf{y}$ , where  $|\boldsymbol{\Sigma}^{-1/2}|$  is the **Jacobian**.

Notice that  $\|\mathbf{z}\|^2 = z_1^2 + z_2^2 + \dots + z_k^2$ , and by **Fubini's Theorem**, we have:

$$\int_{\mathbb{R}^k} \varphi_1(x_1) \cdot \varphi_2(x_2) \cdots \varphi_k(x_k) d\mathbf{x} = \left( \int_{\mathbb{R}} \varphi_1(x_1) dx_1 \right) \left( \int_{\mathbb{R}} \varphi_2(x_2) dx_2 \right) \cdots \left( \int_{\mathbb{R}} \varphi_k(x_k) dx_k \right)$$

Also, there is a well-known result that  $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$ . We are all set now.

Combine everything together, we get:

$$\begin{aligned}
 \int_{\mathbb{R}^k} f(\mathbf{x}) \, d\mathbf{x} &= \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2} \|\boldsymbol{\Sigma}^{-1/2} \mathbf{y}\|^2} \, d\mathbf{y} = \int_{\mathbb{R}^k} \frac{1}{\sqrt{(2\pi)^k}} e^{-\frac{1}{2} \sum_{i=1}^k z_i^2} \, d\mathbf{z} \\
 &= \prod_{i=1}^k \left( \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} \, dz_i \right) \\
 &= 1^k \\
 &= 1
 \end{aligned}$$

Hence proved.

## Problem 2: MLE of Multivariate Gaussian

- (a) Show that  $\frac{\partial}{\partial \mathbf{A}} \text{Tr}(\mathbf{AB}) = \mathbf{B}^\top$ .
- (b) Show that  $\mathbf{x}^\top \mathbf{Ax} = \text{Tr}(\mathbf{xx}^\top \mathbf{A})$ .
- (c) Derive the maximum likelihood estimator for the multivariate Gaussian.

**Solution:**

- (a) Given  $\mathbf{M} \in \mathbb{M}_{n \times n}$ , recall that the definition of the **trace** is stated as:

$$\text{Tr}(\mathbf{M}) = \sum_{i=1}^n m_{ii}.$$

Also, for matrices  $\mathbf{A} \in \mathbb{M}_{n \times m}$  and  $\mathbf{B} \in \mathbb{M}_{m \times n}$ , the matrix multiplication gives:

$$(\mathbf{AB})_{ij} = \sum_{k=1}^m a_{ik} b_{kj} \quad \text{Tr}(\mathbf{AB}) = \sum_{i=1}^n \sum_{k=1}^m a_{ik} b_{ki}$$

Since all the entries of  $\mathbf{A}$  are independent, the derivative rule is given as:

$$\frac{\partial}{\partial a_{lm}} (a_{ik} b_{ki}) = \begin{cases} b_{ki} = b_{ml}, & \text{if } l = i \text{ and } m = k. \\ 0, & \text{otherwise} \end{cases}$$

Recall that the matrix derivative is given as:

$$\frac{\partial}{\partial \mathbf{A}} = \begin{pmatrix} \frac{\partial}{\partial a_{11}} & \cdots & \frac{\partial}{\partial a_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{n1}} & \cdots & \frac{\partial}{\partial a_{nm}} \end{pmatrix}$$

Combine everything, we get:

$$\frac{\partial \text{Tr}(\mathbf{AB})}{\partial \mathbf{A}} = \sum_{i=1}^n \sum_{k=1}^m \begin{pmatrix} \frac{\partial}{\partial a_{11}} & \cdots & \frac{\partial}{\partial a_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{n1}} & \cdots & \frac{\partial}{\partial a_{nm}} \end{pmatrix} a_{ik} b_{ki} = \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{m1} \\ b_{12} & b_{22} & \cdots & b_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{mn} \end{pmatrix} = \mathbf{B}^\top$$

Hence proved.

(b) For  $\mathbf{A} \in M_{n \times n}$  and  $\mathbf{x} = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$ , we have:

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \mathbf{x}^\top \begin{pmatrix} -\mathbf{A}_1 \mathbf{x} - \\ -\mathbf{A}_2 \mathbf{x} - \\ \vdots \\ -\mathbf{A}_n \mathbf{x} - \end{pmatrix} = \sum_{i=1}^n x_i \mathbf{A}_i \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j \quad (1)$$

For  $\mathbf{A}_i$  is the  $i$ th row of the matrix  $\mathbf{A}$ ; also:

$$\mathbf{x} \mathbf{x}^\top \mathbf{A} = \mathbf{x} \begin{pmatrix} \mathbf{x}^\top \mathbf{A}'_1 & \mathbf{x}^\top \mathbf{A}'_2 & \cdots & \mathbf{x}^\top \mathbf{A}'_n \\ | & | & & | \end{pmatrix},$$

where  $\mathbf{A}'_j$  is the  $j$ th column of matrix  $\mathbf{A}$ . Therefore,

$$\text{Tr}(\mathbf{x} \mathbf{x}^\top \mathbf{A}) = \sum_{j=1}^n (\mathbf{x} \mathbf{x}^\top \mathbf{A})_{jj} = \sum_{j=1}^n x_j \sum_{i=1}^n x_i a_{ij} = \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j \quad (2)$$

We obtain the same result in (1) and (2), meaning that:

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n x_i a_{ij} x_j = \text{Tr}(\mathbf{x} \mathbf{x}^\top \mathbf{A})$$

Hence proved.

(c) The original likelihood function of  $N$   $k$ -dimensional multivariate Gaussian is:

$$L(\mu, \Sigma) = \prod_{i=1}^N \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1}(\mathbf{x}^{(i)} - \mu)}$$

After taking log, the log-likelihood function is:

$$\ell(\mu, \Sigma) = \sum_{i=1}^N \left( -\frac{k}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) \right).$$

If we take the derivative with respect to  $\mu$ ,

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \sum_{i=1}^N \frac{1}{2} \left( \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) \right)^\top + \frac{1}{2} (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} = \sum_{i=1}^N \frac{1}{2} (\mathbf{x}^{(i)} - \mu)^\top \left( \Sigma^{-1} + (\Sigma^{-1})^\top \right) \\ &= \sum_{i=1}^N (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} \end{aligned}$$

Letting  $\frac{\partial \ell}{\partial \mu} = \mathbf{0}$  in equation (1), we must have  $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)}$ .

Now we take the derivative with respect to  $\Sigma$ :

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{N}{2} \underbrace{\frac{\partial \ln |\Sigma|}{\partial \Sigma}}_{(i)} - \frac{1}{2} \sum_{i=1}^N \underbrace{\frac{\partial}{\partial \Sigma} \left( (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) \right)}_{(ii)}$$

In equation (1), we have two derivatives to deal with:

(i) For this part, we need several lemmas to cover:

**Lemma 1: Jacobi's Formula**

For  $\mathbf{A} \in \mathbb{M}_{n \times n}$  invertible matrix, we have the following identity:

$$\frac{d}{dt} |\mathbf{A}(t)| = \text{tr} \left( (\mathbf{A}^*(t)) \frac{d\mathbf{A}(t)}{dt} \right) = |\mathbf{A}(t)| \cdot \text{tr} \left( \mathbf{A}(t)^{-1} \cdot \frac{d\mathbf{A}(t)}{dt} \right)$$

In special case:

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = \frac{\partial |\mathbf{A}|}{\partial \mathbf{A}_{ij}} = \mathbf{A}_{ji}^* = (\mathbf{A}^*)^\top$$

Where  $\mathbf{A}^*$  be the **adjoint matrix** of  $\mathbf{A}$ .

**Lemma 2: Adjoint Matrix Identity**

For a **invertible** matrix  $\mathbf{A} \in \mathbb{M}_{n \times n}$ , and  $\mathbf{A}^*$  be  $\mathbf{A}$ 's adjoint, we have:

$$\mathbf{A}^* = |\mathbf{A}| \cdot \mathbf{A}^{-1}$$

With the assist of **lemma 1, 2**, we have:

$$\frac{\partial \ln |\Sigma|}{\partial \Sigma} = \frac{\partial \ln |\Sigma|}{\partial |\Sigma|} \cdot \frac{\partial |\Sigma|}{\partial \Sigma} = \frac{1}{|\Sigma|} \cdot |\Sigma| \cdot \Sigma^{-1} = \Sigma^{-1}$$

(ii) We use the result in (b), we get:

$$\frac{\partial}{\partial \Sigma} \left( (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) \right) = \frac{\partial}{\partial \Sigma} \text{Tr} \left( (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} \right)$$

Recall that  $\text{Tr}(\cdot, \cdot)$  is **reflexive**, *i.e.*  $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$ :

$$\frac{\partial}{\partial \Sigma} \text{Tr} \left( (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} \right) = \frac{\partial}{\partial \Sigma} \text{Tr} \left( \Sigma^{-1} (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top \right)$$

**Lemma 3: Inverse Matrix Derivative**

For  $\mathbf{A} \in \mathbb{M}_{n \times n}$ , and  $\mathbf{A}^{-1}$  be its inverse:

$$d(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}(d\mathbf{A})\mathbf{A}^{-1} \iff d\mathbf{A} = -\mathbf{A}d(\mathbf{A}^{-1})\mathbf{A}$$

With **lemma 3**, we get the following:

$$\begin{aligned} \partial \text{Tr}(\Sigma^{-1} (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top) &= \text{Tr}(\partial(\Sigma^{-1}) (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top) \\ &= \text{Tr}(-\Sigma^{-1}(\partial \Sigma) \Sigma^{-1} (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top) \\ &= \text{Tr}(-(\partial \Sigma) \Sigma^{-1} (\mathbf{x}^{(i)} - \mu)(\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1}) \end{aligned}$$

Therefore, with the result in (a),

$$\begin{aligned}\frac{\partial}{\partial \Sigma} \text{Tr} \left( \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) (\mathbf{x}^{(i)} - \mu)^\top \right) &= - \left( \Sigma^{-1} (\mathbf{x}^{(i)} - \mu) (\mathbf{x}^{(i)} - \mu)^\top \Sigma^{-1} \right)^\top \\ &= - \Sigma^{-1} (\mathbf{x}^{(i)} - \mu)^\top (\mathbf{x}^{(i)} - \mu) \Sigma^{-1}\end{aligned}$$

Bring everything back to the original identity:

$$\frac{\partial \ell}{\partial \Sigma} = -\frac{N}{2} \underbrace{\Sigma^{-1}}_{(i)} + \frac{1}{2} \sum_{i=1}^N \underbrace{\Sigma^{-1} (\mathbf{x}^{(i)} - \mu)^\top (\mathbf{x}^{(i)} - \mu) \Sigma^{-1}}_{(ii)}$$

By letting  $\frac{\partial \ell}{\partial \Sigma} = 0$ , we multiply the equations on the left and right by  $\Sigma$ :

$$0 = -N \hat{\Sigma} + \sum_{i=1}^N (\mathbf{x}^{(i)} - \mu)^\top (\mathbf{x}^{(i)} - \mu) \iff \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}^{(i)} - \mu)^\top (\mathbf{x}^{(i)} - \mu)$$

Finally, we get our maximum likelihood estimators are:

$$\begin{cases} \hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} \\ \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}^{(i)} - \mu)^\top (\mathbf{x}^{(i)} - \mu) \end{cases}$$