Notes on "Pattern Recognition and Machine Learning"

Zhen Wang

July 18, 2013

Contents

1 Introduction

1.1 Information Theory

When we observe a specific value of a random variable $X = x$, the information gain is:

Definition 1.1.1.

$$h(x) = -\log_2 \Pr(x) \tag{1}$$

which is a monotonic function of $\Pr(x)$ and reflects the "degree of surprise". For independent random variables X, Y, $\Pr(x, y) = \Pr(x)\Pr(y)$, thus the information gain $h(x, y) = \log_2 \Pr(x, y) = \log_2 \Pr(x) + \log_2 \Pr(y) = h(x) + h(y)$ which satisfies our intuition.

Entropy can be viewed as the average amount of information to transmit a random variable. Thus is defined as the expectation of (1) with respect to the distribution $\Pr(X)$:

Definition 1.1.2.

$$H[X] = -\sum_x \Pr(x) \log_2 \Pr(x) \tag{2}$$

A deeper interpretation as a measure of disorder:

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . [1]

The more evenly a distribution spreads, the higher its entropy is. Besides, the *noiseless coding theorem* states that:

Conclusion 1.1.1. The entropy is a lower bound on the average number of bits needed to transmit the state of a random variable. We can arrive the lower bound by choosing an efficient coding scheme.

The maximum entropy configuration can be found by maximizing the following equation which includes the normalization constraint on the probability distribution:

$$\tilde{H} = -\sum_{i=1}^{M} p_i \ln(p_i) + \lambda(\sum_{i=1}^{M} p_i - 1) \tag{3}$$

where the λ is a Lagrangian multiplier and doesn't need to be determined. To find stationary points, we check its derivatives:

$$\frac{\partial \tilde{H}}{\partial p_i} = -(\ln(p_i) + 1) + \lambda = 0 \text{ for } i = 1, \ldots, M$$

$$\sum_{i=1}^{M} p_i - 1 = 0$$

We have stationary $p_i = \frac{1}{M}, i = 1, \ldots, M$ and its second derivative is:

$$\frac{\partial \tilde{H}}{\partial p_i \partial p_j} = -I_{ij}\frac{1}{p_i} = -I_{ij}M \tag{4}$$

Since $M > 0$, the second derivative is negative definite (diagonal matrix's eigenvalues are elements on its diagonal), thus the stationary point is indeed a maximum.

The entropy for continuous variable can't be elegantly defined and we use the concept—differential entropy:

Definition 1.1.3.

$$H[X] = -\int \Pr(x) \ln \Pr(x) \mathrm{d}x \tag{5}$$

We find the maximum of $H[x]$ using Lagrange multipliers with normalization constraint of probability distribution and contraints over the first and second moments of $\Pr(X)$:

$$\lambda_1(\int_{-\infty}^{\infty} \Pr(x)\mathrm{d}x - 1)$$

$$\lambda_2(\int_{-\infty}^{\infty} x \Pr(x)\mathrm{d}x - \mu)$$

$$\lambda_3(\int_{-\infty}^{\infty} (x-\mu)^2 \Pr(x)\mathrm{d}x - \sigma^2)$$

Using the calculus of variations and set the derivative of the functional to zero giving us:

.....

.....

.....

.....

.....

..... [2]

$$\Pr(X) = \exp\{-1 + \lambda_1 + \lambda_2 X + \lambda_3 (X-\mu)^2\} \tag{6}$$

back substitution of (6) into the three constraint equations gives us:

.....

.....

.....

.....

.....

.....[3]

$$\Pr(X) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\{-\frac{(X-\mu)^2}{2\sigma^2}\} \tag{7}$$

so we have:

Conclusion 1.1.2. The maximum entropy configuration for continuous random variable is the Guassian.

Then we evaluate its differential entropy. $\ln \Pr(X) = \frac{-1}{2}\ln(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}$ and $-\int_{-\infty}^{\infty}\Pr(X)\frac{-1}{2}\ln(2\pi\sigma^2)\mathrm{d}X = \frac{1}{2}\ln(2\pi\sigma^2)$. So the difficulty is to calculate

$$\begin{aligned}&\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\{\frac{-(X-\mu)^2}{2\sigma^2}\}\frac{(X-\mu)^2}{2\sigma^2}\mathrm{d}X\\ =&\frac{(2\sigma^2)^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1}{2}}}\int_{-\infty}^{\infty}\exp\{\frac{-(X-\mu)^2}{2\sigma^2}\}\frac{(X-\mu)^2}{2\sigma^2}\mathrm{d}\frac{(X-\mu)}{(2\sigma^2)^{\frac{1}{2}}}\end{aligned} \tag{8}$$

By subsection integral method, we know that:

$$\begin{aligned}\int_{-\infty}^{\infty}x^2\exp(-x^2)\mathrm{d}x =& \int_{-\infty}^{\infty}\frac{\mathrm{d}\exp(-x^2)}{\mathrm{d}x}\frac{-x}{2}\mathrm{d}x\\ =&\exp(-x^2)\frac{-x}{2}|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\exp(-x^2)\frac{\mathrm{d}\frac{-x}{2}}{\mathrm{d}x}\mathrm{d}x\\ =&0+\frac{1}{2}\int_{-\infty}^{\infty}\exp(-x^2)\mathrm{d}x\\ =&\frac{\pi^{\frac{1}{2}}}{2}\end{aligned} \tag{9}$$

Thus, based on (9) and (8), we have $H[X] = \frac{1}{2}\{1+\ln(2\pi\sigma^2)\}$.

Suppose we have a joint distribution $\Pr(X,Y)$. If we observe a specific value $X = x$, then the additional information needed to specify the corresponding value of $Y = y$ is given by $-\ln\Pr(y|x)$. Intuitively, $h(x,y) = h(x)+h(y|x)$, by the production rule of probability, $-\ln\Pr(x,y) = -\ln\{\Pr(x)\Pr(y|x)\} = -\ln\Pr(x) - \ln\Pr(y|x)$, which comfirms that our definition makes sense.

Then we define the *conditional entropy* of Y given X as the average additional information needed to specify Y:

Definition 1.1.4.

$$H[Y|X] = -\int\int\Pr(X,Y)\ln\Pr(Y|X)\mathrm{d}Y\mathrm{d}X \tag{10}$$

Then we confirm that $H[X,Y] = H[Y|X] + H[X]$:

$$\begin{aligned}H[X,Y] =& -\int\int\Pr(X,Y)\ln\Pr(X,Y)\mathrm{d}X\mathrm{d}Y\\ =&-\int\int\Pr(X,Y)\ln\Pr(Y|X)+\Pr(X,Y)\ln\Pr(X)\mathrm{d}X\mathrm{d}Y\\ =&H[Y|X] - \int\ln\Pr(X)\mathrm{d}X\int\Pr(X,Y)\mathrm{d}Y\\ =&H[Y|X] - \int\Pr(X)\ln\Pr(X)\mathrm{d}X\\ =&H[Y|X]+H[X]\end{aligned} \tag{11}$$

1.1.1 Relative entropy and mutual information

Definition 1.1.5. *Convex* functions are functions that satisfy the following inequality:

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \tag{12}$$

where $0 \leq \lambda \leq 1$.

A convex function satisfies the *Jensen's inequality*:

$$f(\sum_{i=1}^{M} \lambda_i x_i) \leq \sum_{i=1}^{M} \lambda_i f(x_i) \tag{13}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{M} \lambda_i = 1$. We can prove the above inequality by induction:

Proof.

$$\sum_{i=1}^{N+1} \lambda_i f(x_i) = (1-\lambda_{N+1}) \sum_{i=1}^{N} \frac{\lambda_i}{1-\lambda_{N+1}} f(x_i) + \lambda_{N+1} f(x_{N+1}) \tag{14}$$

$\because$ (13) is valid for previous N cases $\therefore$ (14) $\geq (1-\lambda_{N+1}) f(\sum_{i=1}^{N} \frac{\lambda_i}{1-\lambda_{N+1}} x_i) + \lambda_{N+1} f(x_{N+1})$
$\because$ $f(\cdot)$ is a convex function $\therefore$ we have:

$$(1-\lambda_{N+1}) f(\sum_{i=1}^{N} \frac{\lambda_i}{1-\lambda_{N+1}} x_i) + \lambda_{N+1} f(x_{N+1}) \geq f(\sum_{i=1}^{N+1} \lambda_i x_i) \tag{15}$$

Thus, we derive the $N+1$ case from previous case. □

We can interpret λ_i as the probability over $X = x_i$, then (13) can be written as:

$$f(E[X]) \leq E[f(X)] \tag{16}$$

which is also valid for continuous case.

Suppose a random variable X is transmitted and its distribution $p(X)$ is unknown. We use distribution $q(X)$ to approximate it. Thus the average *additional* amount of information required to specify $X = x$ is:

Definition 1.1.6.

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(X) \ln q(X) \mathrm{d}X - (-\int p(X) \ln p(X) \mathrm{d}X) \\ &= -\int p(X) \ln \{\frac{q(X)}{p(X)}\} \mathrm{d}X \end{aligned} \tag{17}$$

Obviously, *Kullback-Leibler divergence* is not symmetric. Since $(-\ln x)'' = \frac{1}{x^2} > 0$, $-\ln x$ is a convex function. Thus we can apply Jensen inequality to it:

$$\int -\ln \{\frac{q(X)}{p(X)}\} p(X) \mathrm{d}X \geq -\ln \int p(X) \frac{q(X)}{p(X)} \mathrm{d}X = 0 \tag{18}$$

Obviously, $\mathrm{KL}(p\|q) = 0$ only when $q(X) = p(X)$. Besides, we have:

Conclusion 1.1.3. KL divergence is the *lower bound* for average additional information that must be transmitted. We can arrive the bound by choosing an efficient coding scheme.

Both data compression and density estimation are aimed at modelling an unknown probability distribution. Suppose $q(X)$ is characterized by a bunch of parameters θ and we estimate θ by minimizing KL divergence. However, $p(X)$ is unknown, we only have its samples $x_i, i = 1, \ldots, N$ and we use this traning set to approximate $p(X)$:

$$\mathrm{KL}(p\|q) \simeq \sum_{i=1}^{N} \{-\ln q(x_i|\theta) + \ln p(x_i)\} \tag{19}$$

Note that the second term on the right-hand side of (19) is indepedent of θ, and the first term is the negative log likelihood function for θ. Thus we have:

Conclusion 1.1.4. Minimizing KL divergence $\Leftrightarrow$ Maximizing the likelihood function.

Given two random variables X, Y, they are independent if $\Pr(X, Y) = \Pr(X)\Pr(Y)$. We can measure how independent they are by:

Definition 1.1.7.

$$\begin{aligned} I[X, Y] &\equiv \mathrm{KL}(\Pr(X, Y) \| \Pr(X)\Pr(Y)) \\ &= -\int\int \Pr(X, Y) \ln\left(\frac{\Pr(X)\Pr(Y)}{\Pr(X, Y)}\right) \mathrm{d}X\mathrm{d}Y \end{aligned} \tag{20}$$

which is called the *mutual information* between X and Y.

Mutual information is related to conditional entropy:

$$\begin{aligned} I[X, Y] &= -\int\int \Pr(X, Y) \ln\left(\frac{\Pr(X)\Pr(Y)}{\Pr(X, Y)}\right) \mathrm{d}X\mathrm{d}Y \\ &= H[X] + H[Y] - \left(-\int\int \Pr(X, Y) \ln \Pr(X, Y) \mathrm{d}X\mathrm{d}Y\right) \\ &= H[X] + H[Y] - H[X, Y] \\ &= H[X] - H[X|Y] = H[Y] - H[Y|X] \quad \text{(by (11))} \end{aligned} \tag{21}$$

Thus, the mutual information can also be viewed as the residue in the uncertainty about X after being told the value of Y (or vice versa).

2 Probability Distribution

2.1 The Gaussian Distribution

Preliminaries:

Conclusion 2.1.1. For quadratic form $\mathbf{x}^{\mathrm{T}} A\mathbf{x}$, matrix A can be taken to be symmetric without loss of generality.

Proof. $\because \mathbf{x}^{\mathrm{T}} A\mathbf{x} = \sum_{i=1}^{D}\sum_{j=1}^{D} x_i x_j A_{ij} \quad \therefore \forall$ subscript pair (i, j) its contribution is $x_i x_j (A_{ij} + A_{ji})$
Thus, for any asymmetric matrix B, it is equivalent to A where $A_{ij} = A_{ji} = \frac{B_{ij} + B_{ji}}{2}$ □

Conclusion 2.1.2. Matrix A is symmetric $\Leftrightarrow A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y}$.

Proof. $(\Rightarrow) \because A\mathbf{x} \cdot \mathbf{y} = (A\mathbf{x})^{\mathrm{T}}\mathbf{y} = \mathbf{x}^{\mathrm{T}} A\mathbf{y} = \mathbf{x} \cdot A\mathbf{y}$
$(\Leftarrow) A\mathbf{e}_i \cdot \mathbf{e}_j = A_{ji} = \mathbf{e}_i \cdot A\mathbf{e}_j = A_{ij}$
$\because \forall (i, j) : A_{ij} = A_{ji} \quad \therefore A = A^{\mathrm{T}}$ □

Conclusion 2.1.3. Real symmetric matrix $A_{n\times n}$ has n real eigenvalues (counting multiplicities).

Proof. Suppose $A\mathbf{x} = \lambda\mathbf{x}$ where $\mathbf{x} \in \mathbb{C}^n$. Let $q = \bar{\mathbf{x}}^{\mathrm{T}} A\mathbf{x}$
$\because \bar{q} = \mathbf{x}^{\mathrm{T}} A\bar{\mathbf{x}} = \bar{q}^{\mathrm{T}} = \bar{\mathbf{x}}^{\mathrm{T}} A\mathbf{x} = q \quad \therefore q$ is actually real.
$\because q = \bar{\mathbf{x}}^{\mathrm{T}} A\mathbf{x} = \bar{\mathbf{x}}^{\mathrm{T}} \lambda\mathbf{x} = \lambda\|\mathbf{x}\|^2 \quad \therefore \lambda$ is also real. □

Conclusion 2.1.4. Symmetric matrix $A_{n\times n}$ has n eigenvectors forming an orthonormal set.

Proof. (*Assumption*) for $i = 1, \ldots, n-1 \quad (n > 1)$, symmetric matrix $A_{i\times i}$ has i orthonormal eigenvectors and $A_{n\times n}$ is symmetric but doesn't have n orthonormal eigenvectors.

Pick λ_1, one of $A_{n\times n}$'s n eigenvalues and pick a corresponding unit eigenvector $\mathbf{u}_1$ (i.e., $A\mathbf{u}_1 = \lambda_1\mathbf{u}_1$). Let $W = \mathrm{Nul}\ \mathbf{u}_1^{\mathrm{T}}$ be the null space of $\mathbf{u}_1$. Obviously, it is a $n-1$ dimensional subspace of $\mathbb{R}^n$. Then we could choose the basis for W consisting of $n-1$ orthonormal vectors $\mathbb{B} = \{\mathbf{u}_2 \ldots, \mathbf{u}_n\}$ and let $P_{\mathbb{B}} = [\mathbf{u}_2, \ldots, \mathbf{u}_n]$. Although $P_{\mathbb{B}}$ is not square, its rank is $n-1$ and thus gives an one-to-one and onto mapping from $\mathbb{R}^{n-1}$ to W which implies the existence of its *left inverse* $P_{\mathbb{B}}^{-1}$.

$$\begin{array}{ccc} W = \mathrm{Nul}\ \mathbf{u_1}^{\mathrm{T}} & & \Re^{n-1} \\ \mathbf{w} & \underset{P_{\mathcal{B}}}{\overset{P_{\mathcal{B}}^{-1}}{\rightleftharpoons}} & [\mathbf{w}]_{\mathcal{B}} \end{array}$$

Then we prove that multiplication by A defines a linear transformation $T : W \to W$:
$\because A$ is symmetric. $\therefore \forall \mathbf{w} \in W : \mathbf{u}_1 \cdot A\mathbf{w} = A\mathbf{u}_1 \cdot \mathbf{w} = \lambda_1 \mathbf{u}_1 \cdot \mathbf{w} = 0$
Note that the codomain of mapping T is W but the range may not.

$$\begin{array}{ccc} W = \{\mathbf{w} \in \Re^n | \mathbf{u_1}^{\mathrm{T}}\mathbf{w} = 0\} & \xrightarrow{T : A} & V = \{A\mathbf{w} | \mathbf{w} \in W\} \subseteq W \\ \downarrow \quad P_{\mathcal{B}}^{-1} & & \downarrow \quad P_{\mathcal{B}}^{-1} \\ \Re^{n-1} = \{[\mathbf{w}]_{\mathcal{B}} | \mathbf{w} \in W\} & \xrightarrow{T' : M} & M[\mathbf{w}]_{\mathcal{B}} = [A\mathbf{w}]_{\mathcal{B}} \in \Re^{n-1} \end{array}$$

Now our task is to prove the existence of matrix M in the above relationships:
Suppose $\mathbf{w} = r_2\mathbf{u}_2 + \ldots + r_n\mathbf{u}_n$ and since T is a linear transformation which preserves addition and scalar multiplication, we have:

$$T(\mathbf{w}) = r_2 T(\mathbf{u}_2) + \ldots + r_n T(\mathbf{u}_n) \tag{22}$$

Rewrite (22) in basis $\mathbb{B}$ we have:

$$[T(\mathbf{w})]_{\mathbb{B}} = r_2[T(\mathbf{u}_2)]_{\mathbb{B}} + \ldots + r_n[T(\mathbf{u}_n)]_{\mathbb{B}} \tag{23}$$

Note that $[\mathbf{w}]_{\mathbb{B}} = [r_2 \quad \cdots \quad r_n]^{\mathrm{T}}$ and thus:

$$[T(\mathbf{w})]_{\mathbb{B}} = M[\mathbf{w}]_{\mathbb{B}} \tag{24}$$

where $M = [[T(\mathbf{u}_2)]_{\mathbb{B}} \quad \cdots \quad [T(\mathbf{u}_n)]_{\mathbb{B}}]$

M is related to A in terms of eigenvalues and eigenvectors:
Suppose $M[\mathbf{x}]_{\mathbb{B}} = \lambda[\mathbf{x}]_{\mathbb{B}}$, according to (24), $[A\mathbf{x}]_{\mathbb{B}} = \lambda[\mathbf{x}]_{\mathbb{B}}$. $\because [\cdot]_{\mathbb{B}}$ is actually multiplication by $P_{\mathbb{B}}^{-1} : W \to \mathbb{R}^{n-1}$ which is a linear transformation with one-to-one and onto properties. $\therefore \mathbf{x}$ and $[\mathbf{x}]_{\mathbb{B}}$ possess uniquely correspondence and $\mathbf{x}$ is eigenvector of A. Besides, since we choose the inverse mapping $P_{\mathbb{B}}$ to be constructed by basis of W, $\mathbf{x} \in W$.

If there is an orthonormal basis of $\mathbb{R}^{n-1}$ consisting of $n-1$ eigenvectors of $M : [\mathbf{x}_2]_{\mathbb{B}}, \ldots, [\mathbf{x}_n]_{\mathbb{B}}$, then $\mathbf{x}_2, \ldots, \mathbf{x}_n$ are eigenvectors of A. $\because \forall i = 2, \ldots, n : \mathbf{x}_i \in W \quad \therefore \mathbf{u}_1 \cdot \mathbf{x}_i = 0$. Now the remaining task is to prove $\mathbf{x}_2, \ldots, \mathbf{x}_n$ are orthonormal. Remember that $[\mathbf{x}_2]_{\mathbb{B}}, \ldots, [\mathbf{x}_n]_{\mathbb{B}}$ are orthonormal so that if multiplication by $P_{\mathbb{B}}$ preserves dot products (in turn length), then their correspondences are also orthonormal. So the next steps are:

- multiplication by $P_{\mathbb{B}}$ preserves dot products.
- M has $n-1$ orthonormal eigenvectors.

Preserving dot products means $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{n-1} : \mathbf{x} \cdot \mathbf{y} = (P_{\mathbb{B}}\mathbf{x}) \cdot (P_{\mathbb{B}}\mathbf{y})$. It is straightforward by noticing that $P_{\mathbb{B}}^{\mathrm{T}} P_{\mathbb{B}} = I_{(n-1)\times(n-1)}$. Besides, equality is symmetric and we can say $[\cdot]_{\mathbb{B}}$ preserves dot products (i.e., both $P_{\mathbb{B}}$ and $P_{\mathbb{B}}^{-1}$).

To show that M has $n-1$ orthonormal eigenvectors seems difficult, however, note that by induction, a symmetric matrix $M_{(n-1)\times(n-1)}$ has $n-1$ orthonormal eigenvectors. Thus, we try to prove M is symmetric.
$\because A$ is symmetric. $\therefore A\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot A\mathbf{y}$ $\because []_{\mathbb{B}}$ is an isomorphism and preserves dot products. $\therefore$ according to (24), $M[\mathbf{x}]_{\mathbb{B}} \cdot [\mathbf{y}]_{\mathbb{B}} = [\mathbf{x}]_{\mathbb{B}} \cdot M[\mathbf{y}]_{\mathbb{B}}$ which implies that M is symmetric.

Up to now, we showed that A has n orthonormal eigenvectors $\{\mathbf{u}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ which contradicts our initial assumption. Thus, there doesn't exist such a n and $\forall n$: symmetric matrix $A_{n\times n}$ has n orthonormal eigenvectors. □

Conclusion 2.1.5. Symmetric matrix A is orthonormally diagonalizable, i.e., $A = U\Lambda U^{\mathrm{T}}$ where U is orthogonal (i.e., columns are unit orthogonal vectors and rows are unit orthogonal vectors) and Λ is diagonal matrix.

Proof. We have proved that symmetric matrix $A_{n\times n}$ has n orthonormal eigenvectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ $(A\mathbf{u}_i = \lambda_i \mathbf{u}_i)$. Let $U = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_n]$ and $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$, we show that $A = U\Lambda U^{\mathrm{T}}$: $\because U$ consists of unit orthonormal columns. $\therefore$ right multiplied by $U, AU = [\lambda_1\mathbf{u}_1 \quad \cdots \lambda_n\mathbf{u}_n]$ and $U\Lambda U^{\mathrm{T}}U = U\Lambda = [\lambda_1\mathbf{u}_1 \quad \cdots \lambda_n\mathbf{u}_n]$ □

The multivariate Gaussian distribution takes the form:

Definition 2.1.1.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \tag{25}$$

where $\boldsymbol{\mu}$ is the mean vector, Σ is a $D\times D$ covariance matrix with its determinant $|\Sigma|$.
The functional dependence of the Gaussian on $\mathbf{x}$ is through the quadratic form:

$$\Delta^2=(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{26}$$

The quantity Δ is called the *Mahalanobis distance* from $\boldsymbol{\mu}$ to $\mathbf{x}$.

(Based on above propositions) We can choose Σ^{-1} to be symmetric without loss of generality. Actually, we firstly choose the covariance matrix Σ to be symmetric, then Σ is orthogonal diagonalizable:

$$\begin{aligned}\Sigma&=U\Lambda U^{\mathrm{T}}\\&=\begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_D\end{bmatrix}\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_D\end{bmatrix}\begin{bmatrix}\mathbf{u}_1^{\mathrm{T}}\\ \vdots \\ \mathbf{u}_D^{\mathrm{T}}\end{bmatrix}\\&=\begin{bmatrix}\lambda_1\mathbf{u}_1 & \cdots & \lambda_D\mathbf{u}_D\end{bmatrix}\begin{bmatrix}\mathbf{u}_1^{\mathrm{T}}\\ \vdots \\ \mathbf{u}_D^{\mathrm{T}}\end{bmatrix}\\&=\sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\end{aligned} \tag{27}$$

Because Λ is diagonal matrix, its inverse is $\Lambda^{-1}=\begin{bmatrix}\frac{1}{\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{\lambda_D}\end{bmatrix}$. In addition, U is orthogonal ($UU^{\mathrm{T}}=I$). Thus, we have:

$$\Sigma^{-1}=\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} \tag{28}$$

This equation also implicitly proves that

Conclusion 2.1.6. Inverse of symmetric matrix is also symmetric.

Substitute (28) into (26), the quadratic form becomes:

$$\Delta^2=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \tag{29}$$

where $y_i=\mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$. We can interpret $\{y_i\}$ as a new coordinate system by projecting $(\mathbf{x}-\boldsymbol{\mu})$ to $\{\mathbf{u}_i\}$. The functional dependence of $\mathbf{y}=\begin{bmatrix}y_1 & \cdot & y_D\end{bmatrix}$ on $\mathbf{x}$ is captured by the following equation:

$$\mathbf{y}=U^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}) \tag{30}$$

To be properly normalized, Σ should be *positive definite*. Otherwise, see "Chap12":

.....

.....

.....

.....

.....

.....[4]

Integral by substitution ($\mathbf{x}$ replaced by $\mathbf{y}$) requires calculating *Jacobi determinant* where $J_{ij} = \frac{\partial x_i}{\partial y_j} =$. From (30) we know that $\mathbf{x} = U\mathbf{y} + \boldsymbol{\mu}$ and thus $J_{ij} = \frac{\partial x_i}{\partial y_j} = U^{\mathrm{T}}_{ji}$. Obviously, this implies that $J = U$. Then we have:

$$\begin{aligned} |J|^2 &= |U|^2 \\ &= |U^{\mathrm{T}}||U| \quad (\because |A| = |A^{\mathrm{T}}|) \\ &= |U^{\mathrm{T}}U| \quad (\because |A||B| = |AB|) \\ &= |I| = 1 \end{aligned} \tag{31}$$

Use these properties of determinant in above derivation, we also have:

$$\begin{aligned} |\Sigma| &= |U\Lambda U^{\mathrm{T}}| \\ &= |U||\Lambda||U^{\mathrm{T}}| \\ &= |U^{\mathrm{T}}U||\Lambda| \\ &= \prod_{i=1}^{D} \lambda_i \end{aligned} \tag{32}$$

Up to now, we know the substitution results:

$$\begin{aligned} \Pr(\mathbf{y}) &= \Pr(\mathbf{x})|J| \\ &= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\prod_{i=1}^{D} \lambda_i^{\frac{1}{2}}} \exp\{-\frac{1}{2}\sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}\} \\ &= \prod_{i=1}^{D} \frac{1}{(2\pi\lambda_i)^{\frac{1}{2}}} \exp\{-\frac{y_i^2}{2\lambda_i}\} \end{aligned} \tag{33}$$

From the perspective of $\mathbf{y}$ coordinate system, it is straightforward to show that:

$$\int \Pr(\mathbf{y})\mathrm{d}\mathbf{y} = \prod_{i=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_i)^{\frac{1}{2}}} \exp\{-\frac{y_i^2}{2\lambda_i}\}\mathrm{d}y_i = 1 \tag{34}$$

We now check the moment and second moment of Gaussian. By substitution $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$ we have $\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \int \exp\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\Sigma^{-1}\mathbf{z}\}(\mathbf{z} + \boldsymbol{\mu})\mathrm{d}\mathbf{z}$. Note that $\exp(\cdot)$ is an even function and the integral region is the whole $\mathbb{R}^D$. Thus the term of exponential function multiplied by z will vanish and $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$.

Univariate case we consider $\mathbb{E}[x^2]$, while in the multivariate case (D dimensional) we have D^2 pairs $x_i x_j$. Thus we consider:

$$\begin{aligned} \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \int \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \int \exp\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\Sigma^{-1}\mathbf{z}\}(\mathbf{z} + \boldsymbol{\mu})(\mathbf{z} + \boldsymbol{\mu})^{\mathrm{T}}\mathrm{d}\mathbf{z} \end{aligned} \tag{35}$$

Consider $(\mathbf{z} + \boldsymbol{\mu})(\mathbf{z} + \boldsymbol{\mu})^{\mathrm{T}}$ we firstly find that $\mathbf{z}\boldsymbol{\mu}^{\mathrm{T}}$ and $\boldsymbol{\mu}\mathbf{z}^{\mathrm{T}}$ will vanish because of the same reason leveraged in computing moment of Gaussian above. Secondly, $\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ doesn't contain integral variable $\mathbf{z}$ and thus contributes $\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ to final integral result since $\Pr(\cdot)$ is normalized. Remember that from (30) we have

$\mathbf{z} = U\mathbf{y}$ and substitution results:

$$\begin{aligned}
&\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\int \exp\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\Sigma^{-1}\mathbf{z}\}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathrm{d}\mathbf{z} \\
=&\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\int \exp\{-\sum_{i=1}^{D}\frac{y_i^2}{2\lambda_i}\}(U\mathbf{y})(U\mathbf{y})^{\mathrm{T}}\mathrm{d}\mathbf{y} \\
=&\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\int \exp\{-\sum_{i=1}^{D}\frac{y_i^2}{2\lambda_i}\}(\sum_{j=1}^{D}y_j\mathbf{u}_j)(\sum_{k=1}^{D}y_k\mathbf{u}_k^{\mathrm{T}})\mathrm{d}\mathbf{y} \\
=&\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\sum_{j=1}^{D}\sum_{k=1}^{D}\mathbf{u}_j\mathbf{u}_k^{\mathrm{T}}\int \exp\{-\sum_{i=1}^{D}\frac{y_i^2}{2\lambda_i}\}y_jy_k\mathrm{d}\mathbf{y}
\end{aligned} \tag{36}$$

Since y_jy_k is odd function of y_j, y_k respectively, the term with different j, k will vanish. Besides, $\int \mathcal{N}(x|\mu,\sigma)x^2\mathrm{d}x = \sigma$. Thus, we can further simplify (36) to:

$$\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\sum_{j=1}^{D}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\int \exp\{-\sum_{i=1}^{D}\frac{y_i^2}{2\lambda_i}\}y_j^2\mathrm{d}\mathbf{y} = \sum_{j=1}^{D}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}\lambda_j = \Sigma \tag{37}$$

Based on above analysis and calculation we have:

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \\
&= \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \Sigma
\end{aligned} \tag{38}$$

Because Σ governs the covariance, it is called *covariance matrix* and Σ^{-1} is called the *precision matrix*.

2.1.1 Conditional Gaussian distribution

(preliminaries)

Conclusion 2.1.7.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{bmatrix} \tag{39}$$

where A, B, D, D are respectively $p\times p, p\times q, q\times p, q\times q$ matrices and $M = (A-BD^{-1}C)^{-1}$. Note that D is required to be *invertible*. M is known as the *Schur complement* of the matrix on the left-hand side of (39) w.r.t. the submatrix D.

Proof. Multiply $\begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{bmatrix}$ by $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ and check whether it results $I_{(p+q)\times(p+q)} = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & I_{q\times q} \end{bmatrix}$.

Firstly, $MA - MBD^{-1}C = M(A-BD^{-1}C) = I_{p\times p}$.

Secondly, $MB - MBD^{-1}D = MB - MB = 0$.

Thirdly, $-D^{-1}CMA + D^{-1}C + D^{-1}CMBD^{-1}C = D^{-1}C(-MA+I+MBD^{-1}C) = D^{-1}CM(-A+M^{-1}+BD^{-1}C) = 0$.

Finally, $-D^{-1}CMB + D^{-1}D + D^{-1}CMBD^{-1}D = I_{q\times q}$. □

Suppose $\mathbf{x}$ is a D-dimensional vector with Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)$ and we partition $\mathbf{x}$ into two disjoint subsets $\mathbf{x}_a$ and $\mathbf{x}_b$. We can always put $\mathbf{x}_a$ on the first M components of $\mathbf{x}$ by *permutation* over $\boldsymbol{\mu}$ and Σ without loss of generality.

- $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$
- $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$

- $\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$ and the following properties result from $\Sigma^{\mathrm{T}} = \Sigma$
  - $\Sigma_{aa} = \Sigma_{aa}^{\mathrm{T}}$ and $\Sigma_{bb} = \Sigma_{bb}^{\mathrm{T}}$
  - $\Sigma_{ab}^{\mathrm{T}} = \Sigma_{ba}$
- $\Lambda = \Sigma^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$

Using this partitioning we firstly obtain:

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
=&-\frac{1}{2}\begin{bmatrix}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}} & (\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}\begin{bmatrix}\mathbf{x}_a-\boldsymbol{\mu}_a\\ \mathbf{x}_b-\boldsymbol{\mu}_b\end{bmatrix} \\
=&-\frac{1}{2}\begin{bmatrix}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{aa}+(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\Lambda_{ba} & (\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}+(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\Lambda_{bb}\end{bmatrix}\begin{bmatrix}\mathbf{x}_a-\boldsymbol{\mu}_a\\ \mathbf{x}_b-\boldsymbol{\mu}_b\end{bmatrix} \\
=&-\frac{1}{2}\{(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)+(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\Lambda_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)+(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)+(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\Lambda_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b)\}
\end{aligned} \tag{40}$$

We find that as a function of $\mathbf{x}_a$, this is again a quadratic form.

Conclusion 2.1.8. If two sets of random variables are jointly Gaussian, then the conditional probability distribution of one set conditioned on the other is agian Gaussian.

To determine the mean and covariance of $\Pr(\mathbf{x}_a|\mathbf{x}_b)$, we should firstly note that the exponent in a general $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}\,\Sigma)$ can be written as:

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu} \tag{41}$$

We pick out all terms that are second order in $\mathbf{x}_a$ from (40): $-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\Lambda_{aa}\mathbf{x}_a$. Thus the covariance of $\Pr(\mathbf{x}_a|\mathbf{x}_b)$ is Λ_{aa}^{-1}.

We pick out all terms that are linear in $\mathbf{x}_a$ from (40):

$$\begin{aligned}
&\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\Lambda_{aa}\boldsymbol{\mu}_a + \frac{1}{2}\boldsymbol{\mu}_a^{\mathrm{T}}\Lambda_{aa}\mathbf{x}_a - \frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\Lambda_{ba}\mathbf{x}_a \\
=&\mathbf{x}_a^{\mathrm{T}}\Lambda_{aa}\boldsymbol{\mu}_a - \mathbf{x}_a^{\mathrm{T}}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \quad (\because \mathbf{x}^{\mathrm{T}}A\mathbf{y} = \mathbf{y}^{\mathrm{T}}A^{\mathrm{T}}\mathbf{x} \text{ and } \Lambda_{ba}^{\mathrm{T}} = \Lambda_{ab}) \\
=&\mathbf{x}_a^{\mathrm{T}}\{\Lambda_{aa}\boldsymbol{\mu}_a - \Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\}
\end{aligned} \tag{42}$$

Thus the mean of $\Pr(\mathbf{x}_a|\mathbf{x}_b)$ is $\Lambda_{aa}^{-1}\{\Lambda_{aa}\boldsymbol{\mu}_a - \Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\} = \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)$.

If we can ensure Σ_{bb} to be invertible (how?),

.....

.....

.....

.....

.....

.....[5] we can make use of (39) to express Λ_{aa} as $(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$ and hence:

$$\Sigma_{a|b} = \Lambda_{aa}^{-1} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}) \tag{43}$$

we can also make use of (39) to express Λ_{ab} as $-\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}$ and hence:

$$\begin{aligned}
\mu_{a|b} &= \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \Lambda_{aa}^{-1}\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\mathbf{x}_b-\boldsymbol{\mu}_b)
\end{aligned} \tag{44}$$

2.1.2 Marginal Gaussian distributions

(Preliminary)

Conclusion 2.1.9. Suppose A, B, C, D are matrices of dimension $p \times p, p \times q, q \times p, q \times q$ respectively. Besides, D is *invertible*. Then we have:

$$\det(\begin{bmatrix} A & B \\ C & D \end{bmatrix}) = \det(D)\det(A - BD^{-1}C) \tag{45}$$

Proof. .....

.....

.....

.....

.....

..... [6] □

We calculate $\Pr(\mathbf{x}_a)$ by integrated out $\mathbf{x}_b$:

$$\Pr(\mathbf{x}_a) = \int \Pr(\mathbf{x}_a, \mathbf{x}_b) \mathrm{d}\mathbf{x}_b \tag{46}$$

Because we want to integrate out $\mathbf{x}_b$, we should firstly consider terms that are functional to $\mathbf{x}_b$ in (40). Since b is symmetric to a in (40), we can directly use (41), (43) and (44) to derive:

$$\begin{aligned} & -\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\Lambda_{bb}\mathbf{x}_b + \mathbf{x}_b^{\mathrm{T}}\{\Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\} \\ =& -\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\Lambda_{bb}\mathbf{x}_b + \mathbf{x}_b^{\mathrm{T}}\mathbf{m} \quad (\text{where } \mathbf{m} = \Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)) \\ =& -\frac{1}{2}\mathbf{x}_b^{\mathrm{T}}\Lambda_{bb}\mathbf{x}_b + \mathbf{x}_b^{\mathrm{T}}\Lambda_{bb}(\Lambda_{bb}^{-1}\mathbf{m}) - \frac{1}{2}(\Lambda_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\Lambda_{bb}(\Lambda_{bb}^{-1}\mathbf{m}) + \frac{1}{2}(\Lambda_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\Lambda_{bb}(\Lambda_{bb}^{-1}\mathbf{m}) \\ =& -\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\Lambda_{bb}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\Lambda_{bb}^{-1}\mathbf{m} \end{aligned} \tag{47}$$

Note that $\mathbf{m}$ doesn't contain $\mathbf{x}_b$ and the first term in the right-hand side of (47) is the standard quadratic form. Assuming $\mathbf{x}_a$ is M dimensional, integration results:

$$\int \exp\{-\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\Lambda_{bb}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})\}\mathrm{d}\mathbf{x}_b = (2\pi)^{\frac{D-M}{2}}|\Lambda_{bb}^{-1}|^{\frac{1}{2}} \tag{48}$$

while $\frac{1}{2}\mathbf{m}^{\mathrm{T}}\Lambda_{bb}^{-1}\mathbf{m}$ remains as exponent. We firstly expand it as:

$$\begin{aligned} \frac{1}{2}\mathbf{m}^{\mathrm{T}}\Lambda_{bb}^{-1}\mathbf{m} =& \frac{1}{2}(\Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a))^{\mathrm{T}}\Lambda_{bb}^{-1}(\Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)) \\ =& \frac{1}{2}\{\boldsymbol{\mu}_b^{\mathrm{T}}\Lambda_{bb}\boldsymbol{\mu}_b - \boldsymbol{\mu}_b^{\mathrm{T}}\Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - (\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}\boldsymbol{\mu}_b + (\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)\} \\ =& \frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - (\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}\boldsymbol{\mu}_b + \frac{1}{2}\boldsymbol{\mu}_b^{\mathrm{T}}\Lambda_{bb}\boldsymbol{\mu}_b \end{aligned} \tag{49}$$

Then we consider terms that are not functional to $\mathbf{x}_b$ in (40):

$$-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{aa}(\mathbf{x}_a - \boldsymbol{\mu}_a) + (\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}\Lambda_{ab}\boldsymbol{\mu}_b - \frac{1}{2}\boldsymbol{\mu}_b^{\mathrm{T}}\Lambda_{bb}\boldsymbol{\mu}_b \tag{50}$$

Combining (49) and (50) we get:

$$\exp\{-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^{\mathrm{T}}(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(\mathbf{x}_a - \boldsymbol{\mu}_a)\} \tag{51}$$

Then we can prove it is normalized by transforming original normalization term through (45) and substituting (48) into it:

$$\begin{aligned}
&\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}(2\pi)^{\frac{D-M}{2}}|\Lambda_{bb}^{-1}|^{\frac{1}{2}} \\
=&\frac{1}{(2\pi)^{\frac{M}{2}}}|\Lambda|^{\frac{1}{2}}\frac{1}{|\Lambda_{bb}|^{\frac{1}{2}}} \quad (\because AA^{-1}=I, |A||A^{-1}|=1) \\
=&\frac{1}{(2\pi)^{\frac{M}{2}}}\frac{(|\Lambda_{bb}||\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}|)^{\frac{1}{2}}}{|\Lambda_{bb}|^{\frac{1}{2}}} \\
=&\frac{1}{(2\pi)^{\frac{M}{2}}}\frac{1}{|(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}|^{\frac{1}{2}}}
\end{aligned} \tag{52}$$

Conclusion 2.1.10. If two sets of random variables are jointly Gaussian, then the marginal probability distribution of one set integrated out the other is again Gaussian.

$$\Pr(\mathbf{x}_a) = \int \Pr(\mathbf{x}_a, \mathbf{x}_b)\mathrm{d}\mathbf{x}_b = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa}) \tag{53}$$

2.1.3 Bayes' theorem for Gaussian variables

Up to now, the whole story is given a quadratic form $(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Lambda(\mathbf{x}-\boldsymbol{\mu})$, we could partition $\mathbf{x}, \boldsymbol{\mu}, \Lambda$ with consistent size and transform it into two quadratic forms:

$$-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\Lambda\mathbf{x} = -\frac{1}{2}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m})^{\mathrm{T}}\Lambda_{bb}(\mathbf{x}_b - \Lambda_{bb}^{-1}\mathbf{m}) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})(\mathbf{x}_a-\boldsymbol{\mu}_a) \tag{54}$$

where $\mathbf{m} = \Lambda_{bb}\boldsymbol{\mu}_b - \Lambda_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)$ and thus "μ" for the first term on the right-hand side of (54) is $\boldsymbol{\mu}_b - \Lambda_{bb}^{-1}\Lambda_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)$ which is a linear function to $\mathbf{x}_a$.

Hence, the inverse operation can be applied to $\Pr(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Lambda^{-1})$, $\Pr(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|A\mathbf{x}+\mathbf{b}, L^{-1})$ resulting their joint Gaussian distribution $\Pr(\mathbf{z}) = \Pr(\mathbf{x},\mathbf{y}) = \Pr(\mathbf{y}|\mathbf{x})\Pr(\mathbf{x})$.

Suppose the quadratic form is $(\mathbf{z}-\mathbb{E}[\mathbf{z}])^{\mathrm{T}}R(\mathbf{z}-\mathbb{E}[\mathbf{z}])$ and $R = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix}$, $\mathbb{E}[\mathbf{z}] = \begin{bmatrix} \mathbb{E}[\mathbf{z}]_x \\ \mathbb{E}[\mathbf{z}]_y \end{bmatrix}$, then we calculate them according to (54):

- Obviously, $R_{yy} = L$, $\mathbb{E}[\mathbf{z}]_x = \boldsymbol{\mu}$.
- $\because A\mathbf{x}+\mathbf{b} = \mathbb{E}[\mathbf{z}]_y - R_{yy}^{-1}R_{yx}(\mathbf{x}-\mathbb{E}[\mathbf{z}]_x) \therefore A = -R_{yy}^{-1}R_{yx} = -L^{-1}R_{yx}$, i.e., $R_{yx} = -LA$.
- Now $A\mathbf{x}+\mathbf{b} = \mathbb{E}[\mathbf{z}]_y + A(\mathbf{x}-\boldsymbol{\mu})$, thus, $\mathbb{E}[\mathbf{z}]_y = A\boldsymbol{\mu}+\mathbf{b}$.
- Since R is required to be symmetric, $R_{xy} = R_{yx}^{\mathrm{T}} = -A^{\mathrm{T}}L$.
- $\because R_{xx} - R_{xy}R_{yy}^{-1}R_{yx} = R_{xx} - A^{\mathrm{T}}LA = \Lambda \therefore R_{xx} = \Lambda + A^{\mathrm{T}}LA$.

Use (39), we know that:

$$\begin{aligned}
\Pr(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|\begin{bmatrix} \boldsymbol{\mu} \\ A\boldsymbol{\mu}+\mathbf{b} \end{bmatrix}, \begin{bmatrix} \Lambda + A^{\mathrm{T}}LA & -A^{\mathrm{T}}L \\ -LA & L \end{bmatrix}^{-1}) \\
&= \mathcal{N}(\mathbf{z}|\begin{bmatrix} \boldsymbol{\mu} \\ A\boldsymbol{\mu}+\mathbf{b} \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^{\mathrm{T}} \\ A\Lambda^{-1} & L^{-1}+A\Lambda^{-1}A^{\mathrm{T}} \end{bmatrix})
\end{aligned} \tag{55}$$

2.1.4 Maximum likelihood for the Gaussian

(Preliminary)

Definition 2.1.2. The trace of a square matrix is the sum of the elements on its diagonal, i.e., $\mathrm{Tr}(A_{n\times n}) = \sum_{i=1}^{n} A_{ii}$.

Conclusion 2.1.11.

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA) \tag{56}$$

By which we can generalize to *cyclic* property of the trace operator: $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA)$.

Proof. Suppose A, B is of size $n \times m$ and $m \times n$.

$\forall i \in 1, \ldots, n : (AB)_{ii} = \sum_{k=1}^{m} A_{ik}B_{ki}$ Thus $\mathrm{Tr}(AB) = \sum_{i=1}^{n}\sum_{k=1}^{m} A_{ik}B_{ki}$.

The other, $\mathrm{Tr}(BA) = \sum_{i=1}^{m}\sum_{k=1}^{n} B_{ik}A_{ki}$. Obviously, in these two cases, each element of A is matched to a corresponding element of B in the same way. □

Conclusion 2.1.12.

$$\mathrm{Tr}(\mathbf{x}^{\mathrm{T}}A\mathbf{x}) = \mathrm{Tr}(A\mathbf{x}\mathbf{x}^{\mathrm{T}}) \tag{57}$$

Proof. Obviously, the left-hand side equals $\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbf{x}_i\mathbf{x}_j A_{ij}$.

$\mathbf{x}\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} \mathbf{x}_1\mathbf{x}_1 & \cdots & \mathbf{x}_1\mathbf{x}_M \\ \vdots & \ddots & \vdots \\ \mathbf{x}_M\mathbf{x}_1 & \cdots & \mathbf{x}_M\mathbf{x}_M \end{bmatrix}$ and suppose $A\mathbf{x}\mathbf{x}^{\mathrm{T}} = R$, then $R_{ii} = \sum_{j=1}^{M} A_{ij}\mathbf{x}_j\mathbf{x}_i$. Trace is the sum of elements in diagonal, thus it equals $\sum_{i=1}^{M}\sum_{j=1}^{M} A_{ij}\mathbf{x}_j\mathbf{x}_i$. □

Conclusion 2.1.13.

$$\frac{\partial \mathrm{Tr}(AB)}{\partial x} = \mathrm{Tr}(\frac{\partial A}{\partial x}B) \tag{58}$$

when B is fixed or is not functional to x.

Proof. Just write out the subscripts. □

Conclusion 2.1.14.

$$\frac{\partial}{\partial x}(A^{-1}) = -A^{-1}\frac{\partial A}{\partial x}A^{-1} \tag{59}$$

Proof. $\because 0 = \frac{\partial I}{\partial x} = \frac{\partial AA^{-1}}{\partial x} = \frac{\partial A}{\partial x}A^{-1} + A\frac{\partial A^{-1}}{\partial x}$ □

Conclusion 2.1.15. For a symmetric, strictly positive definite matrix $A_{M\times M}$, we have:

$$\frac{\partial}{\partial x}\ln|A| = \mathrm{Tr}(A^{-1}\frac{\partial A}{\partial x}) \tag{60}$$

Proof. $\because A = A^{\mathrm{T}} \therefore$ by (32) we have $|A| = \prod_{i=1}^{M}\lambda_i$

Then the left-hand side can be expressed as:

$$\sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x} \tag{61}$$

Use (27) and (28), we express $A^{-1}\frac{\partial A}{\partial x}$ as:

$$\begin{aligned} A^{-1}\frac{\partial A}{\partial x} &= \sum_{i=1}^{M}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\sum_{j=1}^{M}\frac{\partial \lambda_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}}{\partial x} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \lambda_j\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}}{\partial x} \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}(\frac{\partial \lambda_j}{\partial x}\mathbf{u}_j\mathbf{u}_j^{\mathrm{T}} + \lambda_j\frac{\partial \mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}}{\partial x}) \\ &= \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} + \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_j\mathbf{u}_j^{\mathrm{T}}}{\partial x} \quad (\because \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij}) \end{aligned} \tag{62}$$

$\because \forall i \in 1, \ldots, M : \mathrm{Tr}(\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}) = \|\mathbf{u}_i\|^2 = 1 \therefore \mathrm{Tr}(\sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\partial x}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}) = \sum_{i=1}^{M}\frac{1}{\lambda_i}\frac{\partial \lambda_i}{\lambda x}\mathrm{Tr}(\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}) =$ (61). Hence, next step is to prove the remaining term in the right-hand side of (62) vanishes.

$$
\begin{aligned}
&\mathrm{Tr}\{\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}(\frac{\partial \mathbf{u}_j}{\partial x}\mathbf{u}_j^{\mathrm{T}}+\mathbf{u}_j\frac{\partial \mathbf{u}_j^{\mathrm{T}}}{\partial x})\} \quad (\because (AB)' = A'B + AB') \\
&= \mathrm{Tr}\{\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_j}{\partial x}\mathbf{u}_j^{\mathrm{T}}\} + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \quad (\because \mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij}) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathrm{Tr}(\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_j}{\partial x}\mathbf{u}_j^{\mathrm{T}}) + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \\
&= \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathrm{Tr}(\mathbf{u}_j^{\mathrm{T}}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_j}{\partial x}) + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \quad (\because \mathrm{Tr}(AB) = \mathrm{Tr}(BA)) \\
&= \mathrm{Tr}(\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\lambda_j}{\lambda_i}\mathbf{u}_j^{\mathrm{T}}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_j}{\partial x}) + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \\
&= \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i^{\mathrm{T}}\frac{\partial \mathbf{u}_i}{\partial x}) + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \\
&= \mathrm{Tr}(\sum_{i=1}^{M}\frac{\partial \mathbf{u}_i}{\partial x}\mathbf{u}_i^{\mathrm{T}}) + \mathrm{Tr}(\sum_{i=1}^{M}\mathbf{u}_i\frac{\partial \mathbf{u}_i^{\mathrm{T}}}{\partial x}) \\
&= \mathrm{Tr}(\sum_{i=1}^{M}\frac{\partial \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}}{\partial x}) = \mathrm{Tr}(\frac{\partial \sum_{i=1}^{M}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}}{\partial x}) = \mathrm{Tr}(\frac{\partial UU^{\mathrm{T}}}{\partial x}) = \mathrm{Tr}(\frac{\partial I}{\partial x}) = 0
\end{aligned}
\tag{63}
$$

□

Given $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where each $\mathbf{x}_i$ is drawn *independently* from a D-dimensional Gaussian distribution. We estimate the parameters of the unknown distribution by maximizing the log likelihood function:

$$
\ln \Pr(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{64}
$$

Firstly, we calculate the derivative of (64) w.r.t $\boldsymbol{\mu}$. Suppose resulting vector is $\mathbf{r}$, $\forall i \in 1, \ldots, D : \mathbf{r}_i = \frac{\partial \ln \Pr(\mathbf{X}|\boldsymbol{\mu},\Sigma)}{\partial \mathbf{r}_i}$. Consider a quadratic form $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^{D}\sum_{j=1}^{D}(\mathbf{x}_i - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_j)\Sigma_{ij}^{-1} = \sum_{i=1}^{D}\sum_{j=1}^{D}(\mathbf{x}_i\mathbf{x}_j - \mathbf{x}_j\boldsymbol{\mu}_i - \mathbf{x}_i\boldsymbol{\mu}_j + \boldsymbol{\mu}_i\boldsymbol{\mu}_j)\Sigma_{ij}^{-1}$.

$$
\begin{aligned}
\because \frac{\partial \Delta^2}{\partial \boldsymbol{\mu}_i} &= 2\sum_{j=1}^{D}(\boldsymbol{\mu}_j - \mathbf{x}_j)\Sigma_{ij}^{-1} \\
\therefore \frac{\partial \Delta^2}{\partial \boldsymbol{\mu}} &= 2\Sigma^{-1}(\boldsymbol{\mu} - \mathbf{x})
\end{aligned}
\tag{65}
$$

According to (65), we know that $\frac{\partial \ln \Pr(\mathbf{X}|\mu,\Sigma)}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N}\Sigma^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$ and set it to zero so that:

$$
\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_n \tag{66}
$$

Then we fix $\boldsymbol{\mu}$ in (64) to $\boldsymbol{\mu}_{\mathrm{ML}}$ and calculate derivative of (64) w.r.t. Σ. Suppose the answer is matrix R

where $R_{ij} = \frac{\partial \ln \Pr(\mathbf{X}|\boldsymbol{\mu},\Sigma)}{\partial \Sigma_{ij}}$. Hence, we can consider the derivative of (64) w.r.t. a certain x:

$$
\begin{aligned}
\frac{\partial \ln \Pr(\mathbf{X}|\boldsymbol{\mu},\Sigma)}{\partial x} =& -\frac{N}{2}\frac{\ln|\Sigma|}{\partial x} - \frac{1}{2}\frac{\partial \sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu})}{\partial x} \\
=& -\frac{N}{2}\mathrm{Tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial x}) - \frac{1}{2}\sum_{n=1}^{N}\frac{\partial \mathrm{Tr}((\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu}))}{\partial x} \quad (\because (60)) \\
=& -\frac{N}{2}\mathrm{Tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial x}) - \frac{1}{2}\sum_{n=1}^{N}\frac{\partial \mathrm{Tr}(\Sigma^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}})}{\partial x} \quad (\because (57)) \\
=& -\frac{N}{2}\mathrm{Tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial x}) - \frac{1}{2}\frac{\partial \mathrm{Tr}(\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}})}{\partial x} \\
=& -\frac{N}{2}\mathrm{Tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial x}) - \frac{1}{2}\mathrm{Tr}(\frac{\partial \Sigma^{-1}}{\partial x}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}) \quad (\because (58)) \\
=& -\frac{N}{2}\mathrm{Tr}(\Sigma^{-1}\frac{\partial \Sigma}{\partial x}) - \frac{1}{2}\mathrm{Tr}(-\Sigma^{-1}\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}) \quad (\because (59)) \\
=& -\frac{N}{2}\mathrm{Tr}(\frac{\partial \Sigma}{\partial x}\Sigma^{-1}I) + \frac{1}{2}\mathrm{Tr}(\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}) \quad (\because (56)) \\
=& -\frac{N}{2}\mathrm{Tr}(\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\Sigma\Sigma^{-1}) + \frac{1}{2}\mathrm{Tr}(\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}) \\
=& \frac{1}{2}\mathrm{Tr}(\frac{\partial \Sigma}{\partial x}\Sigma^{-1}\{\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} - N\Sigma\}\Sigma^{-1})
\end{aligned}
\tag{67}
$$

Thus, we require that $(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} - N\Sigma = 0$, i.e.,

$$
\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}} \tag{68}
$$