Gaussian Distribution, Conditioning, Marginalization, Bayes

Chieh Wu

May 2024

1 Gaussian Distribution

A uni-variate and multi-variate Gaussian distribution can be defined as

$$\underbrace{p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}}_{\textbf{Unit Variate Distribution where } x \textbf{ is a scalar.}} \quad \text{where} \quad x, \mu, \sigma \in \mathbb{R} \tag{1}$$

and

$$\underbrace{p(x) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}}_{\textbf{Unit Variate Distribution where } x \textbf{ is a vector.}} \quad \text{where} \quad x, \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}, |\Sigma| = \text{Det}(\Sigma) \tag{2}$$

2 Conditional Gaussian Distribution

2.1 Quick Summary

1. Given a multi-variate Gaussian distribution $p(x)$ where x is a vector. The goal is to split x into 2 vectors (x_a, x_b) and find $p(x_a|x_b)$.
2. We first write the $p(x_a, x_b)$ in terms of x_a and treat x_b as a constant.
3. Since the conditional is the joint divided by a constant, the conditional takes on the form of the joint in terms of x_a. This tells us that the conditional is also a Gaussian distribution.
4. By matching $p(x_a, x_b = \beta)$ with a Gaussian distribution, we can figure out the mean and covariance matrix of $p(x_a|x_b)$.

2.2 The Process

Given a multi-variate Gaussian distribution where $p(x) = p(x_1, x_2, x_3, ...)$, how would we go about finding the conditional distribution $p(x_1, x_2|x_3, ...)$? In general, we can set of variables being conditions as a vector of random variables $x_a = \begin{bmatrix} x_1 & x_2 & ... \end{bmatrix}^\top$ and the variables that are given as $x_b = \begin{bmatrix} x_3 & x_4 & ... \end{bmatrix}^\top$. This implies that we can rewrite the conditional distribution as

$$p(x_1, x_2|x_3, ...) = p(x_a|x_b) \quad \text{where} \quad x = \begin{bmatrix} x_a \\ x_b \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \end{bmatrix} \quad , \quad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Several Facts about Σ

1. Σ is the covariance matrix.
2. Covariance matrices are **always** symmetric where $\Sigma^\top = \Sigma$.
3. The inverse of the covariance matrix is called the **Precision matrix,** $\Sigma^{-1} = \Lambda$.
4. It is often easier to work with Precision matrices when mathematical manipulations are required.
5. The precision matrix can also be split into 4 quadrants like the covariance matrix where

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} = \Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}. \tag{3}$$

6. Since The covariance matrix is symmetric, we know that $\Sigma_{ab} = \Sigma_{ba}$ and $\Lambda_{ab} = \Lambda_{ba}$.

Given these facts, we can rewrite Eq. (2) as

$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \frac{1}{(2\pi)^{d/2}|\Lambda^{-1}|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Lambda (x-\mu)\right\}. \tag{4}$$

For reasons that will become obvious later, we are going to set the constant in front of the exponential term simply as γ. Combining γ with how x, μ, Σ are defined in Eq. (3), it gives us the equation

$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \gamma \exp\left\{-\frac{1}{2}\left(\begin{bmatrix} x_a \\ x_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}\right)^\top \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \left(\begin{bmatrix} x_a \\ x_b \end{bmatrix} - \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}\right)\right\}. \tag{5}$$

To further simplify the notations, we are going to denote

$$\bar{x}_a = x_a - \mu_a \quad \text{and} \quad \bar{x}_b = x_b - \mu_b,$$

to further simplify Eq. (6) into

$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \gamma \exp\left\{\underbrace{-\frac{1}{2}\begin{bmatrix} \bar{x}_a & \bar{x}_b \end{bmatrix} \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \begin{bmatrix} \bar{x}_a \\ \bar{x}_b \end{bmatrix}}_{\textbf{Let's focus on this term as } Q.}\right\}. \tag{6}$$

If we multiply Q out, we would get

$$Q = -\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa} \bar{x}_a + \underbrace{\bar{x}_a^\top \Lambda_{ab} \bar{x}_b + \bar{x}_b^\top \Lambda_{ba} \bar{x}_a}_{\textbf{Pay special attention to these 2 terms}} + \bar{x}_b^\top \Lambda_{bb} \bar{x}_b\right) \tag{7}$$

We purposely set the last term in red because it is the only term that didn't have x_a. Remember, our goal is to go from $p(x_a, x_b)$ to $p(x_a|x_b)$. Therefore, the final result should be in terms of x_a, and everything else can be considered as a constant. Therefore, since the last term didn't include x_a, it can be treated as a constant.

Moreover, Realize that all the terms are scalars. Therefore, the transpose of a scalar is equivalent to its original value. That is,

$$\bar{x}_a^\top \Lambda_{ab} \bar{x}_b = (\bar{x}_a^\top \Lambda_{ab} \bar{x}_b)^\top = \bar{x}_b^\top \Lambda_{ab}^\top \bar{x}_a. \tag{8}$$

Also, from property 6, we also know that $\Lambda_{ab}^\top = \Lambda_{ba}$, therefore

$$\bar{x}_b^\top \Lambda_{ab}^\top \bar{x}_a = \bar{x}_b^\top \Lambda_{ba} \bar{x}_a.$$

This observation leads us to the conclusion that the 2 middle terms of Q must be equivalent, simplifying Q into

$$Q = -\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa} \bar{x}_a + \underbrace{2\bar{x}_a^\top \Lambda_{ab} \bar{x}_b}_{\textit{merged}} + \bar{x}_b^\top \Lambda_{bb} \bar{x}_b\right) \tag{9}$$

2.3 Key realization at this point!

Once we have simplified the joint distribution $p(x_a, x_b)$, we must realize a very important relationship between $p(x_a, x_b)$ and $p(x_a|x_b)$. Given Baye's rule, we know that

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)}.$$

Here, remember that both x_a and x_b are vectors. Therefore, if we are given the vector $x_b = \beta$, we can plug β into both $p(x_a, x_b = \beta)$ and $p(x_b = \beta)$. Let's take a second and think about the consequence of plugging β into these 2 functions.

1. For the joint distribution $p(x_a, x_b = \beta)$ results in the original joint distribution but with $x_b = \beta$ values plugged in.

2. For the marginal distribution $p(x_b = \beta)$, this results in a scalar value for the probability of $p(x_b = \beta)$.

3. The conditional distribution is therefore a distribution where the joint (with β plugged in) divided by some number

$$p(x_a|x_b) = \frac{p(x_a, x_b = \beta)}{\textbf{some number}}.$$

4. We previously simplified the joint distribution as

$$p(x_a, x_b) = \gamma \exp\left\{-\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b + \bar{x}_b^\top \Lambda_{bb}\bar{x}_b\right)\right\} \tag{10}$$

5. Following this logic, the conditional must then be

$$p(x_a|x_b) = \frac{\gamma \exp\left\{-\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b + \bar{x}_b^\top \Lambda_{bb}\bar{x}_b\right)\right\}}{\textbf{some number}} \tag{11}$$

We can split the red constant term out as just another constant

$$p(x_a|x_b) = \frac{\gamma\; e^{\left\{-\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b\right)\right\}} e^{-\frac{1}{2}\bar{x}_b^\top \Lambda_{bb}\bar{x}_b}}{\textbf{some number}} \tag{12}$$

6. We can now combine all the constants together and just call it λ, resulting in

$$p(x_a|x_b) = \lambda\; e^{\left\{-\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b\right)\right\}} \tag{13}$$

7. Key: The posterior distribution looks very similar to a multivariate Gaussian Distribution in terms of x_a. Indeed, with a little more manipulation, the posterior turns out to be another Gaussian distribution.

8. There is a huge advantage in "knowing" that the posterior is a Gaussian distribution. **Knowing the mean and the covariance matrix uniquely identifies the entire distribution.**

9. Therefore, we don't need to use Bayes theorem to calculate the posterior, we simply need to find the mean and the covariance matrix.

10. In the upcoming section, we can see how to get the exact Gaussian distribution

2.4 Further Simplifying the Exponent

We last concluded that the posterior distribution could be written as

$$p(x_a|x_b) = \lambda\; e^{\left\{-\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b\right)\right\}}. \tag{14}$$

Let's further simplify the exponential by writing out the full version.

$$Q = -\frac{1}{2}\left(\bar{x}_a^\top \Lambda_{aa}\bar{x}_a + 2\bar{x}_a^\top \Lambda_{ab}\bar{x}_b\right) = -\frac{1}{2}\left[\underbrace{(x_a^\top - \mu_a^\top)\Lambda_{aa}(x_a - \mu_a)}_{\textbf{1st term}} + \underbrace{2(x_a^\top - \mu_a^\top)\Lambda_{ab}(x_b - \mu_b)}_{\textbf{2nd term}}\right] \tag{15}$$

Let's now multiply the terms out. The constant terms without x_a will again be highlighted in red.

$$Q = -\frac{1}{2}\left(\underbrace{x_a^\top \Lambda_{aa}x_a - \mu_a^\top \Lambda_{aa}x_a - x_a^\top \Lambda_{aa}\mu_a + \mu_a^\top \Lambda_{aa}\mu_a}_{\textbf{1st term}} + \underbrace{2(x_b^\top \Lambda_{ba}x_a - \mu_b^\top \Lambda_{ba}x_a - x_b^\top \Lambda_{ba}\mu_a + \mu_b^\top \Lambda_{ba}\mu_a)}_{\textbf{2nd term}}\right) \tag{16}$$

Here, we once again use the property that scalar terms are equal its transpose. This allows us to put all x_a terms on the left side and simplify.

$$\begin{aligned} Q &= -\frac{1}{2}\left(\underbrace{x_a^\top \Lambda_{aa}x_a - x_a^\top \Lambda_{aa}\mu_a - x_a^\top \Lambda_{aa}\mu_a + \mu_a^\top \Lambda_{aa}\mu_a}_{\textbf{1st term}} + \underbrace{2(x_a^\top \Lambda_{ab}x_b - x_a^\top \Lambda_{ab}\mu_b - x_b^\top \Lambda_{ba}\mu_a + \mu_b^\top \Lambda_{ba}\mu_a)}_{\textbf{2nd term}}\right) && (17) \\ &= -\frac{1}{2}\left(\underbrace{x_a^\top \Lambda_{aa}x_a}_{\textbf{Quadratic Term}} \quad -2x_a^\top \Lambda_{aa}\mu_a + 2x_a^\top \Lambda_{ab}x_b - 2x_a^\top \Lambda_{ab}\mu_b + \textit{constant}\right) && (18) \\ &= \underbrace{-\frac{1}{2}x_a^\top \Lambda_{aa}x_a}_{\textbf{Quadratic Term}} + \underbrace{x_a^\top(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b))}_{\textbf{linear term}} + \textit{constant} && (19) \end{aligned}$$

Therefore, we now know that the conditional distribution **must** look something like the following given some constant c

$$p(x_a|x_b) = c \exp\left\{ -\frac{1}{2}x_a^\top \Lambda_{aa} x_a + x_a^\top \underbrace{(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b))}_{\textbf{Pay special attention here}} + constant \right\}. \tag{20}$$

Let's pay special attention to this equation for later usage. Next, we know that $p(x_a|x_b)$ must be a Gaussian distribution in terms of x_a with mean of $\bar{\mu}$ and precision of $\bar{\Lambda}$. In the form of

$$\begin{aligned}
p(x_a|x_b) &= c \exp\left\{ -\frac{1}{2}\left(x_a^\top - \bar{\mu}\right)^\top \bar{\Lambda}\left(x_a^\top - \bar{\mu}\right) \right\} && (21)\\
&= c \exp\left\{ -\frac{1}{2}\left(x_a^\top \bar{\Lambda} x_a - 2x_a^\top \bar{\Lambda}\bar{\mu} + \underbrace{\bar{\mu}^\top \Lambda \bar{\mu}}_{\textbf{constant}} \right) \right\} && (22)\\
&= c \exp\left\{ -\frac{1}{2}x_a^\top \bar{\Lambda} x_a + x_a^\top \underbrace{\bar{\Lambda}\bar{\mu}}_{\textbf{blue in}\, Eq.\ (20)} + \underbrace{-\frac{1}{2}\bar{\mu}^\top \Lambda \bar{\mu}}_{\textbf{constant}} \right\} && (23)
\end{aligned}$$

By comparing $p(x_a|x_b)$ from the joint distribution in Eq. (20) and the standard Gaussian form in Eq. (23), we come to 2 conclusions.

1. We know from the quadratic term, the precision matrix for $p(x_a|x_b)$ where $\bar{\Lambda} = \Lambda_{aa}$.
2. We also know from the linear term that

$$\bar{\Lambda}\bar{\mu} = \Lambda_{aa}\bar{\mu} = \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b) \tag{24}$$

This give us an expression to solve for $\bar{\mu}$ with

$$\begin{aligned}
\Lambda_{aa}\bar{\mu} &= \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b) && (25)\\
\bar{\mu} &= \Lambda_{aa}^{-1}\left(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b)\right) && (26)\\
&= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b + \mu_b) && (27)
\end{aligned}$$

From this observation, see that the posterior is simply a Gaussian distribution where

$$\mathcal{N}(\bar{\mu}, \bar{\Lambda}) \quad \text{where} \quad \begin{cases} \bar{\mu} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b + \mu_b) \\ \bar{\Lambda} = \Lambda_{aa} \end{cases}. \tag{28}$$

2.5 Using Shur Complement as an Alternative

We have $p(x_a|x_b)$ from Eq. (28), but they are in terms of the precision matrix, Λ. If the joint distribution was originally given as

$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \frac{1}{(2\pi)^{d/2}|\Lambda^{-1}|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Lambda (x-\mu) \right\}, \tag{29}$$

then Eq. (28) tells us directly the posterior $p(x_a|x_b)$. However, if the joint distribution was given as

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu) \right\}, \tag{30}$$

then, we need to take the inverse of Σ to obtain Λ before we can get the posterior. It turns out that there is a trick called ***Shur Complement*** to get $p(x_a|x_b)$ directly even if we started off with Σ. According to Schur Complement, the inverse of a block matrix has the following property.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & (D^{-1} + D^{-1}CMBD^{-1}) \end{bmatrix} \quad \text{where} \quad M = (A - BD^{-1}C)^{-1}. \tag{31}$$

Looking closely at the definition of the covariance matrix and the precision matrix, we can define Λ_{aa} and Λ_{ab} in terms of Σ blocks. Note that

$$\begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}. \tag{32}$$

This implies that Λ_{aa} is equivalent to the M matrix, and $\Lambda_{ab} = -D^{-1}CM$, tellings us that

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \tag{33}$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \tag{34}$$

Since $\bar{\mu}$ and $\bar{\Lambda}$ from Eq. (28) are in terms of $\Lambda_{aa}, \Lambda_{ab}$, we can use it to get $\bar{\mu}$ and $\bar{\Lambda}$ in terms of Σs, giving us

$$\bar{\mu} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \tag{35}$$

$$\bar{\Lambda} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}. \tag{36}$$

3 Marginalization of Gaussian Distributions

3.1 Quick Summary

1. We first write the $p(x_a, x_b)$ in terms of x_b. The x_b portion turns out to be a Gaussian.
2. When we take the integral of a Gaussian, it becomes 1, leaving us the remaining portion of $p(x_a)$.

$$p(x_a) = \int p(x_a, x_b)\ dx_b. \tag{37}$$

3.2 The Detailed Steps

Given a joint Gaussian Distribution, we previously learned to perform conditioning. In this section, we will learn how to marginalize some variables. More specifically, we have a Gaussian distribution

$$\underbrace{p(x) = \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}}_{\textbf{Unit Variate Distribution where } x \textbf{ is a vector.}} \quad \text{where} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_x \end{bmatrix} \tag{38}$$

Similar to what we did in the conditional portion, we also separate the x_i variables into x_a and x_b. Our goal for marginalization is to find $p(x_a)$ where

$$p(x_a) = \int p(x_a, x_b)\ dx_b. \tag{39}$$

It turns out that after your marginalize, $p(x_a)$ is also a Gaussian distribution.

Multivariate Gaussian Bayesian Parameter Estimation

Consider a set of n observations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ drawn from a multivariate Gaussian distribution with unknown mean $\boldsymbol{\mu}$ and known covariance matrix $\boldsymbol{\Sigma}$. We aim to estimate the mean vector $\boldsymbol{\mu}$ using Bayesian inference.

Prior Distribution

We assume a Gaussian prior for the mean vector $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}), \tag{40}$$

where $\boldsymbol{\mu}_0$ is the prior mean and $\boldsymbol{\Lambda}_0$ is the prior covariance matrix.

Likelihood

The likelihood of the observed data $\mathbf{X}$ given the mean vector $\boldsymbol{\mu}$ is:

$$\begin{aligned} p(\mathbf{X}|\boldsymbol{\mu}) &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda^{-1}}) && (41)\\ &= \prod_{i=1}^{n} \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right) \right) && (42)\\ &= \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right), && (43) \end{aligned}$$

where d is the dimensionality of the observations.

Posterior Distribution

The posterior distribution $p(\boldsymbol{\mu}|\mathbf{X})$ is proportional to the product of the prior and the likelihood:

$$\begin{aligned} p(\boldsymbol{\mu}|\mathbf{X}) &\propto p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}) && (44)\\ &= \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right) && (45)\\ &\quad \times \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{1/2}} \right) \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right). && (46) \end{aligned}$$

Combining the exponent terms:

$$-\frac{1}{2} \left[(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right]. \tag{47}$$

Simplifying the Exponent

Expanding and combining the terms inside the exponent given $\boldsymbol{\bar{x}_i} = \frac{1}{n}\sum_i x_i$ and $\boldsymbol{n\bar{x}} = \sum_i x_i$:

$$\begin{aligned} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) &= \sum_{i=1}^{n} \left[\mathbf{x}_i^\top \boldsymbol{\Lambda} \mathbf{x}_i - 2\boldsymbol{\mu}^\top \Lambda \mathbf{x}_i + \boldsymbol{\mu}^\top \boldsymbol{\Lambda} \boldsymbol{\mu} \right] && (48)\\ &= \underbrace{\left(\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Lambda} \mathbf{x}_i \right)}_{\epsilon_1} - 2\boldsymbol{n\mu}^\top \boldsymbol{\Lambda} \bar{\mathbf{x}} + n\boldsymbol{\mu}^\top \boldsymbol{\Lambda} \boldsymbol{\mu} && (49)\\ &= \boldsymbol{n\mu}^\top \boldsymbol{\Lambda} \boldsymbol{\mu} - 2\boldsymbol{n\mu}^\top \boldsymbol{\Lambda} \bar{\mathbf{x}} + \epsilon_1. && (50) \end{aligned}$$

Now, let's also look at the prior term exponent:

$$(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = \boldsymbol{\mu}^\top \Lambda_0 \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \underbrace{\boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0}_{\epsilon_2} \tag{51}$$

Let's now combine the likelihood and posterior

$$-\frac{1}{2}\left(n\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} - 2n\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\bar{\mathbf{x}} + \epsilon_1 + \boldsymbol{\mu}^\top \boldsymbol{\Lambda}_{\mathbf{0}}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \epsilon_2\right) \tag{52}$$

$$-\frac{1}{2}\left(\boldsymbol{\mu}^\top \underbrace{(n\boldsymbol{\Lambda} + \boldsymbol{\Lambda}_{\mathbf{0}})}_{\bar{\boldsymbol{\Lambda}}}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \underbrace{(n\boldsymbol{\Lambda}\bar{\mathbf{x}} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0)}_{\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}} + \underbrace{\epsilon_3}_{\epsilon_1+\epsilon_2}\right) \tag{53}$$

The posterior is in terms of $\boldsymbol{\mu}$ looks like

$$\boldsymbol{\mu}^\top \bar{\boldsymbol{\Lambda}}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}} + \bar{\boldsymbol{\mu}}^\top \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}} \tag{54}$$

where $\bar{\boldsymbol{\Lambda}}$ and $\bar{\boldsymbol{\mu}}$ are the posterior covariance and mean. Looking at Eq. (54), we can look at Eq. (53) and match the terms, telling us that

$$\bar{\boldsymbol{\Lambda}} = n\boldsymbol{\Lambda} + \boldsymbol{\Lambda}_{\mathbf{0}} \tag{55}$$

$$\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}} = n\boldsymbol{\Lambda}\bar{\mathbf{x}} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 \tag{56}$$

$$\bar{\boldsymbol{\mu}} = n\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\Lambda}\bar{\mathbf{x}} + \bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0. \tag{57}$$

We now know the mean and covariance matrix for the posterior distribution.