Gaussian Distribution, Conditioning, Marginalization, Bayes

Chieh Wu

May 2024

Contents

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

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

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

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

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.

$$Q = -\frac{1}{2}\left(\underbrace{x_a^\top \Lambda_{aa}x_a - x_a^\top \Lambda_{aa}\mu_a - x_a^\top \Lambda_{aa}\mu_a + \mu_a^\top \Lambda_{aa}\mu_a}_{\textbf{1st term}} + \underbrace{2(x_a^\top \Lambda_{ab}x_b - x_a^\top \Lambda_{ab}\mu_b - x_b^\top \Lambda_{ba}\mu_a + \mu_b^\top \Lambda_{ba}\mu_a)}_{\textbf{2nd term}}\right) \tag{17}$$

$$= -\frac{1}{2}\left(\underbrace{x_a^\top \Lambda_{aa}x_a}_{\textbf{Quadratic Term}} \quad -2x_a^\top \Lambda_{aa}\mu_a + 2x_a^\top \Lambda_{ab}x_b - 2x_a^\top \Lambda_{ab}\mu_b + constant\right) \tag{18}$$

$$= \underbrace{-\frac{1}{2}x_a^\top \Lambda_{aa}x_a}_{\textbf{Quadratic Term}} + \underbrace{x_a^\top(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b + \mu_b))}_{\textbf{linear term}} + constant \tag{19}$$

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

$$p(x_a|x_b) = c\exp\left\{-\frac{1}{2}\left(x_a^\top - \bar{\mu}\right)^\top \bar{\Lambda}\left(x_a^\top - \bar{\mu}\right)\right\} \tag{21}$$

$$= c\exp\left\{-\frac{1}{2}\left(x_a^\top \bar{\Lambda}x_a - 2x_a^\top \bar{\Lambda}\bar{\mu} + \underbrace{\bar{\mu}^\top \Lambda\bar{\mu}}_{\textbf{constant}}\right)\right\} \tag{22}$$

$$= c\exp\left\{-\frac{1}{2}x_a^\top \bar{\Lambda}x_a + x_a^\top \underbrace{\bar{\Lambda}\bar{\mu}}_{\textbf{blue in } Eq.\ (20)} + \underbrace{-\frac{1}{2}\bar{\mu}^\top \Lambda\bar{\mu}}_{\textbf{constant}}\right\} \tag{23}$$

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

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

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 marginalization, $p(x_a)$ is also a Gaussian distribution.

3.3 Marginalization of a Multivariate Gaussian

Let $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$ follow a multivariate Gaussian distribution:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \;\middle|\; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$. We aim to find $p(\mathbf{x}_a)$ by marginalizing out $\mathbf{x}_b$.

3.4 Step 1: Quadratic Form in the Exponent

The exponent of the Gaussian is:

$$-\frac{1}{2}\begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}^\top \boldsymbol{\Sigma}^{-1} \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}.$$

3.5 Step 2: Block Matrix Inversion

The inverse of the partitioned covariance matrix $\boldsymbol{\Sigma}$ is computed using the **Schur complement**. Let $\mathbf{M} = \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}$. Then:

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa}^{-1} + \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1} & -\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1} & \mathbf{M}^{-1} \end{bmatrix}.$$

3.6 Step 3: Expanding the Quadratic Form (Detailed)

Substitute $\boldsymbol{\Sigma}^{-1}$ into the quadratic form. Let $\Delta_a = \mathbf{x}_a - \boldsymbol{\mu}_a$ and $\Delta_b = \mathbf{x}_b - \boldsymbol{\mu}_b$. Expanding term-by-term:

$$\begin{aligned} \text{Term 1 (Top-left block):} \quad & \Delta_a^\top \left(\boldsymbol{\Sigma}_{aa}^{-1} + \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\right)\Delta_a, \\ \text{Term 2 (Cross terms):} \quad & -\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1}\Delta_b - \Delta_b^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\Delta_a, \\ \text{Term 3 (Bottom-right block):} \quad & \Delta_b^\top \mathbf{M}^{-1}\Delta_b. \end{aligned}$$

Combine all terms:

$$-\frac{1}{2}\Big[\underbrace{\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\Delta_a}_{\text{Term A}} + \underbrace{\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\Delta_a}_{\text{Term B}} - \underbrace{2\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\mathbf{M}^{-1}\Delta_b}_{\text{Term C}} + \underbrace{\Delta_b^\top \mathbf{M}^{-1}\Delta_b}_{\text{Term D}}\Big].$$

3.7 Completing the Square for $\mathbf{x}_b$

Introduce $\boldsymbol{\mu}_{b|a} = \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\Delta_a$. Notice that:

$$\Delta_b - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\Delta_a = \mathbf{x}_b - \boldsymbol{\mu}_{b|a}.$$

Rewrite Term C + Term D as:

$$-\frac{1}{2}\left[(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})^\top \mathbf{M}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})\right].$$

Terms A and B simplify to:

$$-\frac{1}{2}\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\Delta_a.$$

Thus, the exponent becomes:

$$-\frac{1}{2}\left[(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})^\top \mathbf{M}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_{b|a}) + \Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\Delta_a\right].$$

3.8 Step 4: Integrating Out $\mathbf{x}_b$ (Detailed)

The marginalization requires integrating over $\mathbf{x}_b$:

$$p(\mathbf{x}_a) \propto \int \exp\left(-\frac{1}{2}\left[(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})^\top \mathbf{M}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_{b|a}) + \Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\Delta_a\right]\right) d\mathbf{x}_b.$$

The integral over $\mathbf{x}_b$ is a Gaussian integral:

$$\int \exp\left(-\frac{1}{2}(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})^\top \mathbf{M}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_{b|a})\right) d\mathbf{x}_b = (2\pi)^{m/2}|\mathbf{M}|^{1/2},$$

where $m = \dim(\mathbf{x}_b)$. Substituting this back:

$$p(\mathbf{x}_a) \propto (2\pi)^{m/2}|\mathbf{M}|^{1/2}\exp\left(-\frac{1}{2}\Delta_a^\top \boldsymbol{\Sigma}_{aa}^{-1}\Delta_a\right).$$

3.9 Step 5: Simplifying the Marginal Distribution

The original normalization factor $(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}$ combines with the integral result. Using the determinant identity for block matrices:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{aa}| \cdot |\mathbf{M}|,$$

we get:

$$p(\mathbf{x}_a) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}_{aa}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a)\right),$$

where $k = \dim(\mathbf{x}_a)$. Thus:

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_a \;\middle|\; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right).$$

Multivariate Gaussian Bayesian Parameter Estimation given Σ

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:

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{41}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right) \right) \tag{42}$$

$$= \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right), \tag{43}$$

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:

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \tag{44}$$

$$= \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right) \tag{45}$$

$$\times \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{1/2}} \right) \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \right). \tag{46}$$

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 $\bar{\boldsymbol{x}}_{\boldsymbol{i}} = \frac{1}{n}\sum_i x_i$ and $\boldsymbol{n}\bar{\boldsymbol{x}} = \sum_i x_i$:

$$\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) = \sum_{i=1}^{n} \left[\mathbf{x}_i^\top \boldsymbol{\Lambda}\mathbf{x}_i - 2\boldsymbol{\mu}^\top \Lambda \mathbf{x}_i + \boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \right] \tag{48}$$

$$= \underbrace{\left(\sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\Lambda}\mathbf{x}_i \right)}_{\epsilon_1} - 2\boldsymbol{n}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\bar{\mathbf{x}} + n\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \tag{49}$$

$$= \boldsymbol{n}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} - 2\boldsymbol{n}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\bar{\mathbf{x}} + \epsilon_1. \tag{50}$$

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

$$-\frac{1}{2}\left(n\boldsymbol{\mu}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} - 2n\boldsymbol{\mu}^\top\boldsymbol{\Lambda}\bar{\mathbf{x}} + \epsilon_1 + \boldsymbol{\mu}^\top\boldsymbol{\Lambda_0}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + \epsilon_2\right) \tag{52}$$

$$-\frac{1}{2}\left(\boldsymbol{\mu}^\top\underbrace{\left(n\boldsymbol{\Lambda} + \boldsymbol{\Lambda_0}\right)}_{\bar{\boldsymbol{\Lambda}}}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\underbrace{\left(n\boldsymbol{\Lambda}\bar{\mathbf{x}} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right)}_{\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}} + \underbrace{\epsilon_3}_{\epsilon_1+\epsilon_2}\right) \tag{53}$$

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

$$\bar{\boldsymbol{\Lambda}} = n\boldsymbol{\Lambda} + \boldsymbol{\Lambda_0} \tag{55}$$

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

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

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

Multivariate Gaussian BPE for both μ and Σ

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 μ and unknown covariance matrix σI.

Prior Distribution

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

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}), \tag{58}$$

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:

$$p(\mathbf{X}|\boldsymbol{\mu}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{59}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right) \right) \tag{60}$$

$$= \left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) \right), \tag{61}$$

where d is the dimensionality of the observations.

4 Derivation of the Posterior Distribution for an Isometric Gaussian with Normal-Inverse-Gamma Prior

4.1 The Likelihood

We assume that given μ and σ^2, a single observation $x \in \mathbb{R}^d$ is drawn from an isotropic Gaussian distribution:

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2 I) \tag{62}$$

with the density function:

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x - \mu\|^2\right). \tag{63}$$

4.2 Prior Distribution

We assume a Normal-Inverse-Gamma prior:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2) \tag{64}$$

where:

- $\mu|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2\lambda_0^{-1})$:

$$p(\mu|\sigma^2) = \frac{1}{(2\pi\sigma^2/\lambda_0)^{d/2}} \exp\left(-\frac{\lambda_0}{2\sigma^2}\|\mu - \mu_0\|^2\right). \tag{65}$$

- $\sigma^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0)$:

$$p(\sigma^2) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0-1} \exp\left(-\frac{\beta_0}{\sigma^2}\right). \tag{66}$$

4.3 Posterior Distribution

The posterior is given by:

$$p(\mu, \sigma^2|x) \propto p(x|\mu, \sigma^2)p(\mu|\sigma^2)p(\sigma^2). \tag{67}$$

4.4 Let's get started.

Given an isotropic Gaussian distribution

$$P(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}||x-\mu||^2}$$

The conjugate prior consists of 2 distributions

$$P(\mu|\sigma^2) = \frac{1}{(2\pi\frac{\sigma^2}{\lambda_0})^{d/2}} e^{-\frac{\lambda_0}{2\sigma^2}(\mu-\mu_0)^2}$$

and

$$P(\sigma^2) = \Gamma(\sigma^2|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{(\sigma^2)^{\alpha_0+1}\Gamma(\alpha_0)} e^{-\frac{\beta_0}{\sigma^2}}$$

The joint distribution is

$$P(x, \mu, \sigma) = P(x|\mu, \sigma)P(\mu|\sigma)P(\sigma)$$

Our goal is to use the conjugate priors to find the posterior distribution where

$$P(\mu, \sigma|X) = \frac{P(X, \mu, \sigma)}{P(X)}$$

From the rules of conjugate priors, we know that the posterior is

$$p(\mu, \sigma|X) = p(\mu|\sigma)p(\sigma|X) \tag{68}$$

and that $p(\mu|\sigma)$ as a posterior is the same distribution as the prior with updated parameters, where

$$p(\mu|\sigma^2) = \frac{1}{(2\pi\frac{\sigma^2}{\lambda_1})^{d/2}} e^{-\frac{\lambda_1}{2\sigma^2}(\mu-\mu_1)^2} \tag{69}$$

Notice that compared to the prior $p(\mu|\sigma)$, the parameters μ_0, λ_0 were updated to μ_1, λ_1. Therefore, to find the posterior $p(\mu|\sigma)$, we simply need to identify the new μ_1, λ_1 pair.

Similar idea follows for finding the posterior distribution $p(\sigma|X)$. According to conjugate prior theory the posterior $p(\sigma|X)$ is

$$P(\sigma|X) = \frac{\beta_1^{\alpha_1}}{(\sigma^2)^{\alpha_1+1}\Gamma(\alpha_1)} e^{-\frac{\beta_1}{\sigma^2}} \tag{70}$$

Compare the prior $p(\sigma|X)$ to the posterior, we simply need to update parameters

$$\alpha_0, \beta_0 \rightarrow \alpha_1, \beta_1 \tag{71}$$

and we would then know the entire posterior distribution.

4.5 Identifying the posterior $p(\mu|\sigma)$

To identify the posterior $p(\mu|\sigma)$, we first obtain the joint distribution.

$$p(x, \mu, \sigma) = p(x|\mu, \sigma^2)p(\mu|\sigma^2)p(\sigma^2) = Ge^{-\frac{1}{2\sigma^2}||x-\mu||^2} e^{-\frac{\lambda_0}{2\sigma^2}||\mu-\mu_0||^2} p(\sigma) \tag{72}$$

Note that we collected all the constant terms from $P(x|\mu, \sigma^2)P(\mu|\sigma^2)$ into the term G. This simplifies the terms and allows us to focus mainly on the exponent terms shown in red. Let's now really zoom into the exponent terms

$$-\frac{1}{2\sigma^2}(x-u)^T(x-u) - \frac{\lambda_0}{2\sigma^2}(u-u_0)^T(u-u_0) \tag{73}$$

$$-\frac{1}{2\sigma^2}(x^Tx - 2u^Tx + u^Tu) - \frac{\lambda_0}{2\sigma^2}(u^Tu - 2u_0^Tu + u_0^Tu_0) \tag{74}$$

Our goal is to collect the terms so $p(\mu|\sigma)$ posterior emerges.

Quick Note: As a quick side note, remember that the exponent of the posterior distribution is "supposed" to look like

$$\frac{-1}{2\sigma^2}(\mu^T\lambda_1\mu - 2\mu^T\lambda_1\mu_1 + \mu_1^T\lambda_1\mu_1) \tag{75}$$

Let's reorganize the exponent terms we were working with to match the posterior.

$$-\frac{1}{2\sigma^2}(x^Tx - 2u^Tx + u^Tu) - \frac{\lambda_0}{2\sigma^2}(u^Tu - 2u_0^Tu + u_0^Tu_0) \tag{76}$$

$$-\frac{1}{2\sigma^2}(\mu^T(1+\lambda_0)\mu - 2\mu^T(x+\lambda_0\mu_0) + x^Tx + \lambda_0\mu_0^T\mu_0) \tag{77}$$

and compare what we have versus what it is supposed to look like

$$\underbrace{-\frac{1}{2\sigma^2}(\mu^T\lambda_1\mu - 2\mu^T\lambda_1\mu_1 + \mu_1^T\lambda_1\mu_1)}_{\text{posterior form}}. \tag{78}$$

By matching the terms, we identify λ_1 and μ_1 posterior for $p(\mu|\sigma)$ as

$$\lambda_1 = 1 + \lambda_0 \quad \text{and} \quad \lambda_1\mu_1 = x + \lambda_0\mu_0 \implies \mu_1 = \frac{x + \lambda_0\mu_0}{\lambda_1}. \tag{79}$$

4.6 Finding the Posterior for $p(\sigma|X)$

Now that we have identified λ_1, μ_1, we can substitute these terms back in Eq. (77).

$$-\frac{1}{2\sigma^2}(\mu^T\lambda_1\mu - 2\mu^T\lambda_1\mu_1 + x^Tx + \lambda_0\mu_0^T\mu_0) \tag{80}$$

Again, remember that the posterior $p(\mu|\sigma)$ is supposed to look like

$$-\frac{1}{2\sigma^2}(\mu^T\lambda_1\mu - 2\mu^T\lambda_1\mu_1 + \underbrace{\mu_1^T\lambda_1\mu_1}_{\textbf{this term is currently missing}}) \tag{81}$$

Comparing to what we have against what it is supposed to look like, the term $\mu_1^T\lambda_1\mu_1$ is still missing. Here, we can simply add the term into our equation as a 0. Giving us

$$-\frac{1}{2\sigma^2}(\mu^T\lambda_1\mu - 2\mu^T\lambda_1\mu_1 + \underbrace{\mu_1^T\lambda_1\mu_1 - \mu_1^T\lambda_1\mu_1}_{\textbf{adding 0}} + x^Tx + \lambda_0\mu_0^T\mu_0) \tag{82}$$

By creating the final term, we can now separate out the parts that belong to the posterior $p(\mu|\sigma)$ from the other parts.

$$\underbrace{\frac{-\lambda_1}{2\sigma^2}(\mu-\mu_1)^2}_{\textbf{belong to } p(\mu|\sigma)} - \underbrace{\frac{1}{2\sigma^2}(x^Tx + \lambda_0\mu_0^T\mu_0 - \mu_1^T\Lambda_1\mu_1)}_{\textbf{must be part of } p(\sigma)} \tag{83}$$

Let's now connect and combine the 2nd term with $p(\sigma)$ prior exponent. Giving us

$$\underbrace{-\frac{1}{2\sigma^2}(x^Tx + \lambda_0\mu_0^T\mu_0 - \mu_1^T\Lambda_1\mu_1) - \frac{2\beta_0}{2\sigma^2}}_{\textbf{must all be part of } p(\sigma)}. \tag{84}$$

After a quick simplification, we have

$$-\frac{1}{2\sigma^2}(x^Tx + \lambda_0\mu_0^T\mu_0 - \mu_1^T\Lambda_1\mu_1 + 2\beta_0). \tag{85}$$

We can significantly simplify this term by writing out the term $\mu_1^\top\lambda_1\mu_1$.

Since

$$\mu_1 = \frac{\lambda_0\mu_0 + x}{\lambda_0 + 1}, \quad \lambda_1 = \lambda_0 + 1 \tag{86}$$

Then

$$\mu_1^T\lambda_1\mu_1 = \frac{(\lambda_0\mu_0 + x)^T}{\lambda_0+1}(\lambda_0+1)\frac{(\lambda_0\mu_0+x)}{\lambda_0+1} \tag{87}$$

$$= \frac{1}{\lambda_0+1}(\lambda_0\mu_0 + x)^T(\lambda_0\mu_0 + x) \tag{88}$$

$$= \frac{1}{\lambda_0+1}(\lambda_0^2\mu_0^T\mu_0 + 2x^T\lambda_0\mu_0 + x^Tx). \tag{89}$$

This implies that we have

$$-\frac{1}{2\sigma^2}(x^Tx + \lambda_0\mu_0^T\mu_0 - \mu_1^T\lambda_1\mu_1 + 2\beta_0) \tag{90}$$

where

$$\mu_1^T\lambda_1\mu_1 = \frac{1}{\lambda_1}(\lambda_0^2\mu_0^T\mu_0 + 2x^T\lambda_0\mu_0 + x^Tx) \tag{91}$$

$$= \frac{\lambda_0^2}{\lambda_1}\mu_0^T\mu_0 + 2\frac{\lambda_0}{\lambda_1}x^T\mu_0 + \frac{1}{\lambda_1}x^Tx \tag{92}$$

Plugging $\mu_1^\top\lambda_1\mu_1$ into our equation, we now have

$$-\frac{1}{2\sigma^2}((1-\frac{1}{\lambda_1})x^Tx + (\lambda_0 - \frac{\lambda_0^2}{\lambda_1})\mu_0^T\mu_0 - 2\frac{\lambda_0}{\lambda_1}x^T\mu_0 + 2\beta_0) \tag{93}$$

Note that:

$$\frac{1}{\lambda_1} - \frac{1}{\lambda_1} = \frac{\lambda_0+1-1}{\lambda_1} = \frac{\lambda_0}{\lambda_1} \tag{94}$$

$$\lambda_0 - \frac{\lambda_0^2}{\lambda_1} = \frac{\lambda_0\lambda_1 - \lambda_0^2}{\lambda_1} = \frac{\lambda_0(\lambda_1-\lambda_0)}{\lambda_1} \tag{95}$$

$$= \frac{\lambda_0((\lambda_0+1)-\lambda_0)}{\lambda_1} = \frac{\lambda_0}{\lambda_1} \tag{96}$$

We can use the notes to further simplify our expression.

$$-\frac{1}{2\sigma^2}\left(\frac{1}{\lambda_1}(\lambda_0(x^Tx+\mu_0^T\mu_0-2x^T\mu_0)+2\beta_0)\right) \tag{97}$$

Remember that the exponent term for $p(\sigma|X)$ was

$$-\frac{\beta_1}{\sigma^2} \tag{98}$$

We can move $\frac{1}{2}$ in to make it exactly the same where

$$-\frac{1}{2\sigma^2}\left(\frac{\lambda_0}{\lambda_1}(x-\mu_0)^2+2\beta_0\right) \tag{99}$$

$$-\frac{1}{2\sigma^2}\left(\frac{1}{2}\frac{\lambda_0}{\lambda_1}(x-\mu_0)^2+\beta_0\right) \tag{100}$$

From this, we conclude that

$$\beta_1=\frac{1}{2}(x-\mu_0)^2\frac{\lambda_0}{\lambda_1}+\beta_0 \tag{101}$$

This leaves us with α_1 as the last term to find. Remember that the posterior has a σ^2 term where

$$(\sigma^2)^{-\alpha_1+1} \tag{102}$$

This implies that if we want to find α_1, we simply need to multiply all the σ^2 terms together. They were inside G.
If we collect all the σ^2 terms from $p(x,\mu,\sigma^2)$, we have

$$\frac{1}{(2\pi\sigma^2)^{d/2}}\frac{1}{(2\pi\sigma^2/\lambda_0)^{d/2}}(\sigma^2)^{-\alpha_0-1} \tag{103}$$

Now that we have β_1, that leaves α_1 as the last value we need to find. To find this, we go back to the initial joint distribution.

$$P(x,\mu,\sigma^2)=Qe^{-\frac{\lambda_1}{2\sigma^2}(\mu-\mu_1)^2}e^{-\frac{\beta_1}{\sigma^2}} \tag{104}$$

Notice that the joint distribution is written in terms of the updated values $\Lambda_1,\mu_1,\alpha_1,\beta_1$. We have also incorporated all the constant values from $p(x,\mu,\sigma^2)$ into a single term Q.

In order for us to find α_1, we need to make a slight adjustment to the equation. Namely, we can multiply it by 1.

$$P(x,\mu,\sigma^2)=Q\frac{(2\pi\sigma^2/\lambda_1)^{d/2}}{(2\pi\sigma^2/\lambda_1)^{d/2}}e^{-\frac{\lambda_1}{2\sigma^2}(\mu-\mu_1)^2}e^{-\frac{\beta_1}{\sigma^2}} \tag{105}$$

The reason we multiply by 1 is because it allows us to separate out $p(\mu|\sigma,X)$ posterior completely as

$$p(\mu|\sigma^2)=\frac{1}{(2\pi\sigma^2/\lambda_1)^{d/2}}e^{-\frac{\lambda_1}{2\sigma^2}(\mu-\mu_1)^2} \tag{106}$$

This is really helpful because the posterior is

$$p(\mu|\sigma^2)p(\sigma^2|X)=\underbrace{\frac{1}{(2\pi\sigma^2/\lambda_1)^{d/2}}e^{-\frac{\lambda_1}{2\sigma^2}(\mu-\mu_1)^2}}_{p(\mu|\sigma^2)}\underbrace{Q(2\pi\sigma^2/\lambda_1)^{d/2}e^{-\frac{\beta_1}{\sigma^2}}}_{p(\sigma^2|X)}, \tag{107}$$

implying that everything else must be $p(\sigma^2|X)$.

Now we remove and ignore $p(\mu|\sigma^2)$, leaving us with $p(\sigma^2|X)$ as

$$p(\sigma^2|X)=Q\frac{1}{(2\pi\sigma^2/\lambda_1)^{d/2}}e^{-\frac{\beta_1}{\sigma^2}} \tag{108}$$

Let's now look more carefully into Q.

$$Q=\frac{1}{(2\pi\sigma^2)^{d/2}}\frac{1}{(2\pi\sigma^2/\lambda_0)^{d/2}}\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0-1} \tag{109}$$

Therefore

$$Q(2\pi\sigma^2/\lambda_1)^{d/2} = \frac{(2\pi\sigma^2/\lambda_1)^{d/2}}{(2\pi\sigma^2)^{d/2}} \frac{1}{(2\pi\sigma^2/\lambda_0)^{d/2}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\sigma^2)^{-\alpha_0-1} \tag{110}$$

$$= \left(\frac{\lambda_0}{2\pi\lambda_1}\right)^{d/2} (\sigma^2)^{-(d/2+\alpha_0)-1} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tag{111}$$

Let's now look side by side the posterior we have versus what it is supposed to look like.

What we have

$$p(\sigma^2|X) = \left(\frac{\lambda_0}{(2\pi\lambda_1)}\right)^{d/2} (\sigma^2)^{-(d/2+\alpha_0)-1} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tag{112}$$

and what we want

$$p(\sigma^2|X) = \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} (\sigma^2)^{-\alpha_1-1} e^{-\frac{\beta_1}{\sigma^2}} \tag{113}$$

Focusing on the exponents of the (σ^2) term, notice that α_1 must therefore be

$$\alpha_1 = d/2 + \alpha_0 \tag{114}$$

We now know the posterior $p(\sigma^2|X)$.

5 Normal-Wishart Prior and Posterior Updates

We assume a dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where each $\mathbf{x}_i \in \mathbb{R}^d$ follows a multivariate normal distribution:

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where both the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ are unknown.

5.1 Normal-Wishart Prior

The **Normal-Wishart** prior assumes:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\mathbf{W}_0, \nu_0)$$

where:

- $\boldsymbol{\mu}_0$ is the prior mean.
- κ_0 is the scaling factor controlling the precision of $\boldsymbol{\mu}$.
- $\mathbf{W}_0$ is the prior scale matrix.
- ν_0 is the degrees of freedom.

6 Likelihood Function

The likelihood function for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given N observations is:

$$p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The sufficient statistics of the dataset are:

- **Sample mean**: $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$.
- **Scatter matrix**:

$$\mathbf{S} = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T.$$

7 Posterior Updates

Using the conjugacy property of the Normal-Wishart prior, the posterior remains a **Normal-Wishart distribution**:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}_N, \frac{1}{\kappa_N}\boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma}|\mathcal{D} \sim \mathcal{W}^{-1}(\mathbf{W}_N, \nu_N)$$

where the updated parameters are:

7.1 Updated Posterior Parameters

- **Updated Mean Hyperparameter**:

$$\boldsymbol{\mu}_N = \frac{\kappa_0 \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}}{\kappa_0 + N}$$

- **Updated Scaling Factor**:

$$\kappa_N = \kappa_0 + N$$

- **Updated Degrees of Freedom**:

$$\nu_N = \nu_0 + N$$

- **Updated Scale Matrix**:

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \mathbf{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T$$

8 Understanding the $\mathbf{W}_0$ Matrix

The matrix $\mathbf{W}_0$ in the inverse-Wishart prior is a **scale matrix**, encoding prior information about the covariance structure.

8.1 Wishart and Inverse-Wishart Distributions

The **Wishart distribution**, denoted as:

$$\boldsymbol{\Sigma} \sim \mathcal{W}(\nu, \mathbf{W})$$

has the probability density function:

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}^{-1}\boldsymbol{\Sigma})\right).$$

The **inverse-Wishart distribution**, denoted as:

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\nu, \mathbf{W})$$

has the density:

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{W}\boldsymbol{\Sigma}^{-1})\right).$$

8.2 Role of $\mathbf{W}_0$

- It serves as a prior estimate of the covariance matrix.
- Larger values in $\mathbf{W}_0$ correspond to stronger beliefs about high variances.
- Off-diagonal elements represent prior assumptions about correlations.

8.3 Posterior Update for $\mathbf{W}_0$

After observing data, the posterior scale matrix $\mathbf{W}_N$ is updated as:

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \mathbf{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T.$$

This combines prior knowledge ($\mathbf{W}_0$) with the observed data ($\mathbf{S}$).

9 Summary

The **Normal-Wishart prior**:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\mathbf{W}_0, \nu_0)$$

leads to a **Normal-Wishart posterior**:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}_N, \frac{1}{\kappa_N}\boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma}|\mathcal{D} \sim \mathcal{W}^{-1}(\mathbf{W}_N, \nu_N)$$

with updated parameters:

$$\boldsymbol{\mu}_N = \frac{\kappa_0\boldsymbol{\mu}_0 + N\bar{\mathbf{x}}}{\kappa_0 + N}, \quad \kappa_N = \kappa_0 + N, \quad \nu_N = \nu_0 + N.$$

$$\mathbf{W}_N^{-1} = \mathbf{W}_0^{-1} + \mathbf{S} + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T.$$

This result provides a principled Bayesian update for estimating the mean and covariance of a Gaussian distribution.

10 Bayesian Parameter Estimation

10.1 Using normal Inverse Gamma Prior

In this case, we have data D which we want to fit to a Gaussian distribution. We assume not to know μ and σ^2. In this case, the likelihood function is

$$p(D|\mu,\sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}.$$

The normal Inverse Gamma Prior is

$$p(\mu,\sigma^2) = \mathcal{N}(\mu|\mu_0,\sigma^2V_0)IG(\sigma^2|a_0,b_0). \tag{115}$$

Notice that σ^2 in normal and Inverse Gamma are dependent.

$$\mathcal{N}(\mu|\mu_0,\sigma^2V_0) = (2\pi)^{-1/2}(\sigma^2V_0)^{-1/2}e^{-\frac{1}{2\sigma^2V_0}\sum_i(\mu-\mu_0)^2} \tag{116}$$

$$IG(\sigma^2|a_0,b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1}e^{-b/\sigma^2}. \tag{117}$$

Putting the two equations together, we have the prior for the Gaussian distribution

$$p(\mu,\sigma^2) = \mathcal{N}(\mu|\mu_0,\sigma^2V_0)IG(\sigma^2|a_0,b_0) \tag{118}$$

$$= \underbrace{(2\pi)^{-1/2}(\sigma^2V_0)^{-1/2}e^{-\frac{1}{2\sigma^2V_0}\sum_i(\mu-\mu_0)^2}}_{\mathcal{N}(\mu|\mu_0,\sigma^2V_0)}\underbrace{\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1}e^{-b/\sigma^2}}_{IG(\sigma^2|a_0,b_0)} \tag{119}$$

$$= (2\pi V_0)^{-1/2}\frac{1}{\sigma}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1}e^{-\frac{1}{2\sigma^2V_0}\sum_i(\mu-\mu_0)^2-b/\sigma^2} \tag{120}$$

$$= (2\pi V_0)^{-1/2}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{\sigma}(\sigma^2)^{-a_0-1}e^{-\frac{1}{2\sigma^2}(V_0^{-1}\sum_i(\mu-\mu_0)^2+2b)} \tag{121}$$

10.2 Posterior

We can obtain the posterior distribution by combining the joint likelihood distribution and the prior.

$$p(\mu,\sigma^2|D) = \frac{p(D|\mu,\sigma^2)p(\mu,\sigma^2)}{p(D)}. \tag{122}$$

Since Normal Inverse Gamma distribution is the conjugate prior, the posterior is also Normal Inverse Gamma.

$$p(D|\mu,\sigma^2)p(\mu,\sigma^2) = \underbrace{(2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}}\underbrace{(2\pi V_0)^{-1/2}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{\sigma}(\sigma^2)^{-a_0-1}e^{-\frac{1}{2\sigma^2}(V_0^{-1}(\mu-\mu_0)^2+2b)}} \tag{123}$$

Let's first focus on the exponents by combining them

$$-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2 - \frac{1}{2\sigma^2}(V_0^{-1}\sum_i(\mu-\mu_0)^2+2b) \tag{124}$$

$$-\frac{1}{2\sigma^2}\left(\sum_i(x_i-\mu)^2 + (V_0^{-1}\sum_i(\mu-\mu_0)^2+2b)\right) \tag{125}$$

$$-\frac{1}{2\sigma^2}\left(\sum_i(x_i^2-2x_i\mu+\mu^2) + V_0^{-1}(\mu^2-2\mu\mu_0+\mu_0^2)+2b\right) \tag{126}$$

$$-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - 2(\sum_i x_i)\mu + n\mu^2 + V_0^{-1}\mu^2 - 2V_0^{-1}\mu\mu_0 + V_0^{-1}\mu_0^2 + 2b\right) \tag{127}$$

$$-\frac{1}{2\sigma^2}\left((n+V_0^{-1})\mu^2 + -2(n\bar{X}+V_0^{-1}\mu_0)\mu + V_0^{-1}\mu_0^2 + \sum_i x_i^2 + 2b\right) \tag{128}$$

We know that the posterior is also a Normal Inverse Gamma distribution, therefore, we can look at the exponents of a normal distribution and figure out the mean and the standard deviation. If we assume the posterior has a normal distribution of μ_n, σ_n^2, then the exponents would have the terms

$$-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2 \tag{129}$$

$$-\frac{1}{2\sigma_n^2}\mu^2 + \frac{1}{\sigma_n^2}\mu_n\mu - \frac{1}{2\sigma_n^2}\mu_n^2. \tag{130}$$

If we match the terms of Eq. (128) and Eq. (130), we first see that

$$-\frac{1}{2\sigma_n^2}\mu^2 = -\frac{1}{2\sigma^2}(n+V_0^{-1})\mu^2 \tag{131}$$

implying that we can find the variance of the posterior Gaussian distribution σ_n^2

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma^2}(n+V_0^{-1}) \quad \text{or} \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma^2}V_n^{-1} \quad \text{or} \quad \sigma_n^2 = \frac{\sigma^2}{(n+V_0^{-1})} \tag{132}$$

where we let

$$V_n^{-1} = n + V_0^{-1}. \tag{133}$$

Matching the 2nd terms of Eq. (128) to Eq. (130), we can also find the mean of the posterior distribution μ_n where

$$\frac{1}{\sigma_n^2}\mu_n\mu = \frac{1}{\sigma^2}(n\bar{X} + V_0^{-1}\mu_0)\mu \implies \mu_n = \frac{\sigma_n^2}{\sigma^2}(n\bar{X} + V_0^{-1}\mu_0). \tag{134}$$

Since we previously got an expression of σ_n^2 from Eq. (132), we can plug it in

$$\mu_n = \frac{\sigma^2}{(n+V_0^{-1})}\frac{1}{\sigma^2}(n\bar{X} + V_0^{-1}\mu_0) = \frac{1}{(n+V_0^{-1})}(n\bar{X} + V_0^{-1}\mu_0). \tag{135}$$

We can now rewrite Eq. (128) to emphasize the normal distribution portion of the Normal Inverse Gamma structure as

$$-\frac{1}{2}\underbrace{\frac{(n+V_0^{-1})}{\sigma^2}}_{\frac{1}{\sigma_n^2}}\mu^2 + \underbrace{\frac{1}{\sigma^2 V_n}}_{\frac{1}{\sigma_n^2}}\underbrace{\frac{(n\bar{X}+V_0^{-1}\mu_0)}{(n+V_0^{-1})}}_{\mu_n}\mu - \frac{1}{2}\underbrace{\frac{1}{\sigma^2 V_n}}_{\frac{1}{\sigma_n^2}}\mu_n^2 + \frac{1}{2}\frac{1}{\sigma^2 V_n}\mu_n^2 - \frac{1}{2\sigma^2}V_0^{-1}\mu_0^2 - \frac{1}{2\sigma^2}\sum_i x_i^2 - \frac{1}{2\sigma^2}2b. \tag{136}$$

Once we have identified the normal portion of the equation, we lastly identify the inverse gamma portion of the equation as $-b_n/\sigma^2$, by rewriting the equation into the same structure as

$$-\frac{1}{2}\underbrace{\frac{(n+V_0^{-1})}{\sigma^2}}_{\frac{1}{\sigma_n^2}}\mu^2 + \underbrace{\frac{1}{\sigma^2 V_n}}_{\frac{1}{\sigma_n^2}}\underbrace{\frac{(n\bar{X}+V_0^{-1}\mu_0)}{(n+V_0^{-1})}}_{\mu_n}\mu - \frac{1}{2}\underbrace{\frac{1}{\sigma^2 V_n}}_{\frac{1}{\sigma_n^2}}\mu_n^2 - \frac{b + \frac{1}{2}\left(V_0^{-1}\mu_0^2 + \sum_i x_i^2 - V_n^{-1}\mu_n^2\right)}{\sigma^2}. \tag{137}$$

This tells us that the b_n for the posterior distribution must be

$$b_n = b + \frac{1}{2}\left(V_0^{-1}\mu_0^2 + \sum_i x_i^2 - V_n^{-1}\mu_n^2\right). \tag{138}$$

To identify the posterior distribution,

$$p(\mu_n, \sigma_n^2|D) = \mathcal{N}(\mu|\mu_n, \sigma^2 V_n) IG(\sigma^2|a_n, b_n) \tag{139}$$

$$= (2\pi V_n)^{-1/2}\frac{b_n^{a_n}}{\Gamma(a_n)}\frac{1}{\sigma}(\sigma^2)^{-a_n-1}e^{-\frac{1}{2\sigma^2}(V_n^{-1}\sum_i(\mu-\mu_n)^2+2b)} \tag{140}$$

we need to lastly identify a_n. This can be easily found by identifying the power of the σ^2 term. When we multiply the original terms together, we previously focused on the exponents, but let's now focus on the multipliers.

$$p(D|\mu,\sigma^2)p(\mu,\sigma^2) = \underbrace{(2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}}\underbrace{(2\pi V_0)^{-1/2}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{\sigma}(\sigma^2)^{-a_0-1}e^{-\frac{1}{2\sigma^2}(V_0^{-1}(\mu-\mu_0)^2+2b)}}$$

If we only look at the non-exponent terms, and multiply them together, we have

$$(2\pi)^{-(n+1)/2}(V_0)^{-1/2}\frac{b_0^{a_0}}{\Gamma(a_0)}\frac{1}{\sigma}(\sigma^2)^{(-(n/2+a_0)-1)}. \tag{141}$$

From this, we can conclude that

$$a_n = a_0 + n/2. \tag{142}$$

10.3 Conclusion.

Assuming data D, we assume it is a Gaussian distribution with unknown μ, σ^2. We assume to use Normal-Inverse Gamma distribution. Then the posterior distribution is

$$p(\mu, \sigma^2|D) = \mathcal{N}(\mu|\mu_n, \sigma^2 V_n) IG(\sigma^2|a_n, b_n) \tag{143}$$

$$= (2\pi V_n)^{-1/2} \frac{b_n^{a_n}}{\Gamma(a_n)} \frac{1}{\sigma} (\sigma^2)^{-a_n - 1} e^{-\frac{1}{2\sigma^2}(V_n^{-1} \sum_i (\mu - \mu_n)^2 + 2b_n)} \tag{144}$$

where

$$a_n = a_0 + \frac{n}{2} \tag{145}$$

$$b_n = b_0 + \frac{1}{2}\left[\mu_0^2 V_0^{-1} + \sum_i x_i^2 - \mu_n^2 V_n^{-1}\right] \tag{146}$$

$$V_n^{-1} = V_0^{-1} + n \tag{147}$$

$$\mu_n = \frac{V_0^{-1}\mu_0 + n\bar{X}}{V_0^{-1} + n}. \tag{148}$$

11 Bayesian Linear Regression Notes

11.1 Joint Gaussian Distribution of IID samples

Given x_i, we wish to predict y_i in a regression problem. The Bayesian approach assumes that

$$y_i = x_i^\top w + \epsilon, \tag{149}$$

which implies that y_i has the probability distribution of

$$p(y_i) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(y_i - x_i^\top w)^2}{2\sigma^2}}. \tag{150}$$

Given the distribution of a single sample, then the joint distribution is

$$p(y) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{(y_i - x_i^\top w)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_i^n (y_i - x_i^\top w)^2}. \tag{151}$$

We can convert the exponent portion into a more compact format using the following derivation.

$$\sum_i^n (y_i - x_i^\top w)^2 = \begin{bmatrix}(y_1 - x_1^\top w) & (y_2 - x_2^\top w) & ...\end{bmatrix} \begin{bmatrix}(y_1 - x_1^\top w) \\ (y_2 - x_2^\top w) \\ ...\end{bmatrix} \tag{152}$$

Since

$$\begin{bmatrix}(y_1 - x_1^\top w) \\ (y_2 - x_2^\top w) \\ ...\end{bmatrix} = \left[\begin{bmatrix}y_1 \\ y_2 \\ ...\end{bmatrix} - \begin{bmatrix}x_1^\top \\ x_2^\top \\ ...\end{bmatrix} w\right] = y - Xw, \tag{153}$$

then the exponent can be compactly written as

$$\sum_i^n (y_i - x_i^\top w)^2 = (y - Xw)^\top (y - Xw). \tag{154}$$

The joint distribution could then be compactly written as

$$p(y) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y - Xw)^\top (y - Xw)}. \tag{155}$$

11.2 Conjugate Prior of the Joint Distribution for w

It has already been previously identified that the conjugate prior for w from a Gaussian distribution is simply another Gaussian distribution. Therefore, given $w \in \mathbb{R}^d$, the conjugate prior for w is therefore a d dimensional Gaussian distribution.

$$p(w|\sigma^2, \mu_0, \Sigma_0) = (2\pi\sigma^2)^{-d/2} |\Sigma_0|^{-d/2} e^{-\frac{1}{2\sigma^2}(w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0)}. \tag{156}$$

Note that this distribution assumes that σ^2 is given. Also, often $\Sigma_0 = I$ and $\mu_0 = 0$.

11.3 Conjugate Prior of the Joint Distribution for σ^2

For σ^2 the conjugate prior is the Inverse-Gamma Distribution,

$$p(\sigma^2|a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2}. \tag{157}$$

11.4 Total Joint Distribution

Given the likelihood function and the priors, we have

$$p(y, w, \sigma^2|\mu_0, \Sigma_0, a, b) = p(y|w, \sigma^2) p(w|\sigma^2, \mu_0, \Sigma_0) p(\sigma^2|a, b). \tag{158}$$

Once we have the joint distribution, we can simply sample from w and σ^2 to obtain their distribution. Alternatively, since we used the conjugate prior for the likelihood function, the posterior distribution can be obtained directly by combining all the distributions.

$$p(y, w, \sigma^2|\mu_0, \Sigma_0, a, b) = \underbrace{(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y - Xw)^\top (y - Xw)}}_{p(y)} \underbrace{(2\pi\sigma^2)^{-d/2} |\Sigma_0|^{-d/2} e^{-\frac{1}{2\sigma^2}(w - \mu_0)^\top \Sigma_0^{-1} (w - \mu_0)}}_{p(w|\sigma^2, \mu_0, \Sigma_0)} \underbrace{\frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} e^{-b/\sigma^2}}_{p(\sigma^2|a,b)}$$

To simplify this expression, we first combine all the exponent terms (shown in red).

$$-\frac{1}{2\sigma^2}(y-Xw)^\top(y-Xw)-\frac{1}{2\sigma^2}(w-\mu_0)^\top\Sigma_0^{-1}(w-\mu_0)-b/\sigma^2 \tag{159}$$

$$-\frac{1}{2\sigma^2}(w^\top X^\top Xw-2y^\top Xw+y^\top y)-\frac{1}{2\sigma^2}(w^\top\Sigma_0^{-1}w-2\mu_0^\top\Sigma_0^{-1}w+\mu_0^\top\Sigma_0^{-1}\mu_0)-2b/2\sigma^2 \tag{160}$$

$$-\frac{1}{2\sigma^2}\left[w^\top X^\top Xw-2y^\top Xw+y^\top y+w^\top\Sigma_0^{-1}w-2\mu_0^\top\Sigma_0^{-1}w+\mu_0^\top\Sigma_0^{-1}\mu_0+2b\right] \tag{161}$$

$$-\frac{1}{2\sigma^2}\left[w^\top(X^\top X+\Sigma_0^{-1})w-2(y^\top X+\mu_0^\top\Sigma_0^{-1})w+y^\top y+\mu_0^\top\Sigma_0^{-1}\mu_0+2b\right] \tag{162}$$

11.5 Matching the Exponent Term with Gaussian

We know previously that the Gaussian and Inverse-Gamma distributions are **conjugate priors**, implying that the posterior is also a Gaussian and Inverse-Gamma distribution. Let's rewrite the prior Gaussian distribution here and put it next to the expected posterior Gaussian distribution.

- The prior distribution

$$p(w|\sigma^2,\mu_0,\Sigma_0)=(2\pi\sigma^2)^{-d/2}|\Sigma_0|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu_0)^\top\Sigma_0^{-1}(w-\mu_0)}. \tag{163}$$

- The expected posterior distribution

$$p(w|X,Y,\sigma^2,\mu_0,\Sigma_0)=(2\pi\sigma^2)^{-d/2}|\Sigma|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu)^\top\Sigma^{-1}(w-\mu)}. \tag{164}$$

- Pay special attention that
  - μ_0 and Σ_0 are for the prior distribution
  - μ and Σ are for the posterior distribution once data X and label y are considered.
- To further simplify the notation, we normally represent the Σ^{-1} as the precision matrix Λ, where $\Sigma^{-1}=\Lambda$. Therefore, we can rewrite the posterior as

$$p(w|X,Y,\sigma^2,\mu_0,\Sigma_0)=(2\pi\sigma^2)^{-d/2}|\Lambda^{-1}|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu)^\top\Lambda(w-\mu)}. \tag{165}$$

From Eq. (165), we see that we can automatically identify the entire posterior distribution if we find out μ and Λ. This can be done by matching the various terms from the exponent. If we look more closely at the exponents of Eq. (165), the exponent takes the form of

$$-\frac{1}{2\sigma^2}(w-\mu)^\top\Lambda(w-\mu)=-\frac{1}{2\sigma^2}(w^\top\Lambda w-2\mu^\top\Lambda w+\mu\Lambda\mu). \tag{166}$$

We now have the exponent of the joint distribution as well as the posterior distribution. From the joint distribution, we should be able to identify the posterior distribution. This is easily seen if we put these two equations, Eq. (162) and Eq. (166), side by side.

- The posterior distribution exponent

$$-\frac{1}{2\sigma^2}(w^\top\Lambda w-2\mu^\top\Lambda w+\mu\Lambda\mu) \tag{167}$$

- The joint distribution exponent that is supposed to match the posterior

$$-\frac{1}{2\sigma^2}\left[w^\top(X^\top X+\Sigma_0^{-1})w-2(y^\top X+\mu_0^\top\Sigma_0^{-1})w+y^\top y+\mu_0^\top\Sigma_0^{-1}\mu_0+2b\right] \tag{168}$$

Placing the two equations next to each other, we conclude that the precision matrix matched in red Λ is

$$\Lambda=(X^\top X+\Sigma_0^{-1}). \tag{169}$$

We can also identify μ for the resulting Gaussian distribution. To do this, we match the 2nd term (green terms) and see that

$$\begin{aligned}
2(y^\top X+\mu_0^\top\Sigma_0^{-1})w&=2\mu^\top\Lambda w\\
(y^\top X+\mu_0^\top\Sigma_0^{-1})&=\mu^\top\Lambda\\
(X^\top y+\Sigma_0^{-1}\mu_0)&=\Lambda\mu \qquad note:\Lambda=\Lambda^\top,\Sigma_0^{-1}=(\Sigma_0^{-1})^\top\\
\Lambda^{-1}(X^\top y+\Sigma_0^{-1}\mu_0)&=\mu\\
(X^\top X+\Sigma_0^{-1})^{-1}(X^\top y+\Sigma_0^{-1}\mu_0)&=\mu \qquad given: Eq.\ (169)
\end{aligned}$$

Now that we have defined μ and Λ, we can rewrite Eq. (162) in these terms as

$$-\frac{1}{2\sigma^2}\left[w^\top\Lambda w - 2\mu^\top\Lambda w + \underbrace{\mu^\top\Lambda\mu - \mu^\top\Lambda\mu}_{\textbf{equivalent to 0}} + y^\top y + \mu_0^\top\Sigma_0^{-1}\mu_0 + 2b\right] \tag{170}$$

The reason why we added 0 is because it allows us to separate the exponent of the Gaussian distribution and Inverse Gamma distribution.

$$\underbrace{-\frac{1}{2\sigma^2}(w-\mu)^\top\Lambda(w-\mu)}_{\textbf{Gaussian Distribution Portion}} \quad \underbrace{-\frac{1}{2\sigma^2}(y^\top y + \mu_0^\top\Sigma_0^{-1}\mu_0 + 2b - \mu^\top\Lambda\mu)}_{\textbf{Inverse-Gamma Portion}} \tag{171}$$

11.6 The posterior Gaussian distribution

From the previous derivations, we can conclude that the posterior of w given data is

$$p(w|X,y,\sigma^2,\mu_0,\Sigma_0) = (2\pi\sigma^2)^{-d/2}|\Sigma|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu)^\top\Sigma^{-1}(w-\mu)}. \tag{172}$$

where

$$\Lambda = (X^\top X + \Sigma_0^{-1}) \quad \text{and} \quad \mu = (X^\top X + \Sigma_0^{-1})^{-1}(X^\top y + \Sigma_0^{-1}\mu_0). \tag{173}$$

11.7 The posterior of the Inverse Gamma distribution (IG)

We rewrite the prior Inverse Gamma distribution (IG) equation here as

$$p(\sigma^2|a,b) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b/\sigma^2}. \tag{174}$$

This implies that the posterior distribution would look like

$$p(\sigma^2|X,y,a,b) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}. \tag{175}$$

Pay special attention that we use a, b for the prior and α, β for the posterior. For an IG, we can identify the entire distribution by knowing the α, β parameters. Therefore, we simply need to match up the leftover terms from the Gaussian distribution to identify α and β of the posterior.
From Eq. (171), we rewrite the leftover exponent terms next to the exponent of the IG. This will allow us to easily match the terms.

- Leftover portion for IG

$$-\frac{1}{2\sigma^2}(y^\top y + \mu_0^\top\Sigma_0^{-1}\mu_0 + 2b - \mu^\top\Lambda\mu) \tag{176}$$

$$-\frac{1}{\sigma^2}(\frac{y^\top y}{2} + \frac{\mu_0^\top\Sigma_0^{-1}\mu_0}{2} + b - \frac{\mu^\top\Lambda\mu}{2}) \tag{177}$$

- The posterior distribution

$$-\frac{\beta}{\sigma^2} \tag{178}$$

Having the two equations side by side, we conclude that

$$\beta = \frac{y^\top y}{2} + \frac{\mu_0^\top\Sigma_0^{-1}\mu_0}{2} + b - \frac{\mu^\top\Lambda\mu}{2}. \tag{179}$$

We next need to identify α. We can do this by looking at the power of the σ^2 term from the joint distribution. For convenience, let's rewrite the original joint distribution here as

$$p(y,w,\sigma^2|\mu_0,\Sigma_0,a,b) = \underbrace{(2\pi\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}(y-Xw)^\top(y-Xw)}}_{p(y)}\underbrace{(2\pi\sigma^2)^{-d/2}|\Sigma_0|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu_0)^\top\Sigma_0^{-1}(w-\mu_0)}}_{p(w|\sigma^2,\mu_0,\Sigma_0)}\underbrace{\frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1}e^{-b/\sigma^2}}_{p(\sigma^2|a,b)}$$

Looking specifically at the σ^2 terms shown in red. If we combine them together, we'd get

$$\underbrace{(\sigma^2)^{-d/2}}_{\textbf{Gaussian posterior}} \quad \underbrace{(\sigma^2)^{-n/2-a-1}}_{\textbf{IG posterior}}. \tag{180}$$

We now write the Gaussian posterior next to the IG posterior. You will notice how the σ^2 terms are split between these 2 distributions.

- Gaussian posterior

$$p(w|X,y,\sigma^2,\mu_0,\Sigma_0) = (2\pi\sigma^2)^{-d/2}|\Sigma|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu)^\top\Sigma^{-1}(w-\mu)}. \tag{181}$$

- IG posterior

$$p(\sigma^2|X,y,a,b) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}. \tag{182}$$

Notice that

- $(\sigma^2)^{-d/2}$ belongs to the Gaussian distribution.
- $(\sigma^2)^{-(n/2+a)-1}$ belongs to the IG distribution.
- Implying that

$$\alpha = n/2 + a \tag{183}$$

Once we have identified the posterior IG distribution, we have

$$p(\sigma^2|X,y,a,b) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2} \tag{184}$$

$$\alpha = n/2 + a \tag{185}$$

$$\beta = \frac{y^\top y}{2} + \frac{\mu_0^\top\Sigma_0^{-1}\mu_0}{2} + b - \frac{\mu^\top\Lambda\mu}{2}. \tag{186}$$

11.8 The Predictive Posterior Distribution

Once we have identified the distribution for w and σ^2 as $p(w|X,y,\sigma^2,\mu_0,\Sigma_0)$ and $p(\sigma^2|X,y,a,b)$, we can now make prediction on $\hat{y}$ given $\hat{x}$ as the following distribution.

$$p(\hat{y}|\hat{x},X,y) = \int p(\hat{y},w,\sigma^2|\hat{x},X,y)\,dwd\sigma^2 \tag{187}$$

$$p(\hat{y}|\hat{x},X,y) = \int p(\hat{y}|\hat{x},X,y,w,\sigma^2)p(w|X,y,\sigma^2,\mu_0,\Sigma_0)p(\sigma^2|X,y,a,b)\,dwd\sigma^2. \tag{188}$$

To marginalize the variables, we would combine all 3 distributions together and identify distributions based on w and σ^2 while leaving everything else outside the integral. Since the integral of a distribution is 1, the remainder constant in terms of $\hat{y}$ is the resulting marginalized distribution.
We first identify the 3 distributions as

$$p(\hat{y}|\hat{x},X,y,w,\sigma^2) = (2\pi\sigma^2)^{-1/2}e^{-\frac{(\hat{y}-\hat{x}^\top w)^2}{2\sigma^2}}. \tag{189}$$

$$p(w|X,y,\sigma^2,\mu_0,\Sigma_0) = (2\pi\sigma^2)^{-d/2}|\Sigma|^{-d/2}e^{-\frac{1}{2\sigma^2}(w-\mu)^\top\Sigma^{-1}(w-\mu)} \tag{190}$$

$$p(\sigma^2|X,y,a,b) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2} \tag{191}$$

$$\tag{192}$$

Our first goal is to identify the Gaussian distribution in terms of w. We can do this by combining and simplifying the exponent terms with the following derivation.

$$-\frac{(\hat{y}-\hat{x}^\top w)^\top(\hat{y}-\hat{x}^\top w)}{2\sigma^2} - \frac{1}{2\sigma^2}(w-\mu)^\top\Sigma^{-1}(w-\mu) - \beta/\sigma^2 \tag{193}$$

$$-\frac{1}{2\sigma^2}\left[(\hat{y}-\hat{x}^\top w)^\top(\hat{y}-\hat{x}^\top w) + (w-\mu)^\top\Sigma^{-1}(w-\mu) + 2\beta\right] \tag{194}$$

$$-\frac{1}{2\sigma^2}\left[\underbrace{\hat{y}^\top\hat{y} - 2\hat{y}\hat{x}^\top w + w^\top\hat{x}\hat{x}^\top w} + \underbrace{w^\top\Sigma^{-1}w - 2\mu^\top\Sigma^{-1}w + \mu^\top\Sigma^{-1}\mu} + 2\beta\right] \tag{195}$$

Now we can combine the w terms together.

$$-\frac{1}{2\sigma^2}\left[w^\top(\hat{x}\hat{x}^\top + \Sigma^{-1})w - 2(\hat{y}\hat{x}^\top + \mu^\top\Sigma^{-1})w + \hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu + 2\beta\right] \tag{196}$$

Looking at Eq. (167), it tells us that the Gaussian predictive posterior is supposed to look like

$$-\frac{1}{2\sigma^2}(w^\top\Lambda_p w - 2\mu_p^\top\Lambda_p w + \mu_p\Lambda\mu_p) \tag{197}$$

Note that Λ_p and μ_p are for the predictive posterior. This implies that

$$\Lambda_p = (\hat{x}\hat{x}^\top + \Sigma^{-1}) \tag{198}$$

$$\mu_p = (\hat{x}\hat{x}^\top + \Sigma^{-1})^{-1}(\hat{y}\hat{x}^\top + \mu^\top\Sigma^{-1}) \tag{199}$$

We can now rewrite Eq. (196) as

$$-\frac{1}{2\sigma^2}\left[w^\top\Lambda_p w - 2\mu_p^\top\Lambda_p w + \underbrace{\mu_p^\top\Lambda_p\mu_p - \mu_p^\top\Lambda_p\mu_p}_{=0} + \hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu + 2\beta\right] \tag{200}$$

$$-\frac{1}{2\sigma^2}\left[(w-\mu_p)^\top\Lambda_p(w-\mu_p) - \mu_p^\top\Lambda_p\mu_p + \hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu + 2\beta\right] \tag{201}$$

$$-\frac{1}{2\sigma^2}\left[(w-\mu_p)^\top\Lambda_p(w-\mu_p)\right] - \frac{1}{2\sigma^2}\left[\hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu - \mu_p^\top\Lambda_p\mu_p + 2\beta\right] \tag{202}$$

The joint distribution is therefore,

$$p(\hat{y}, w, \sigma^2|X, y) = (2\pi\sigma^2)^{-1/2}(2\pi\sigma^2)^{-d/2}|\Sigma|^{-d/2}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\frac{1}{2\sigma^2}\left[(w-\mu_p)^\top\Lambda_p(w-\mu_p)\right] - \frac{1}{2\sigma^2}\left[\hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu - \mu_p^\top\Lambda_p\mu_p + 2\beta\right]} \tag{203}$$

Notice there is a Gaussian distribution hidden inside the joint distribution as

$$p(w) = (2\pi\sigma^2)^{-d/2}|\Lambda_p^{-1}|^{-d/2}e^{-\frac{1}{2\sigma^2}\left[(w-\mu_p)^\top\Lambda_p(w-\mu_p)\right]}. \tag{204}$$

This implies that

$$p(\hat{y}, \sigma^2|X, y) = (2\pi\sigma^2)^{-1/2}|\Sigma|^{-d/2}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\frac{1}{2\sigma^2}\left[\hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu - \mu_p^\top\Lambda_p\mu_p + 2\beta\right]}|\Lambda_p^{-1}|^{d/2}\underbrace{\int_w p(w)\, dw}_{==1} \tag{205}$$

$$= (2\pi)^{-1/2}|\Sigma|^{-d/2}|\Lambda_p^{-1}|^{d/2}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-1/2-\alpha-1}e^{-\frac{1}{2\sigma^2}\left[\hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu - \mu_p^\top\Lambda_p\mu_p + 2\beta\right]}. \tag{206}$$

We next marginalize out σ^2 by first setting

$$c = (2\pi)^{-1/2}|\Sigma|^{-d/2}|\Lambda_p^{-1}|^{d/2}\frac{\beta^\alpha}{\Gamma(\alpha)}, \tag{207}$$

$$\beta_p = \frac{1}{2}\left[\hat{y}^\top\hat{y} + \mu^\top\Sigma^{-1}\mu - \mu_p^\top\Lambda_p\mu_p + 2\beta\right] \tag{208}$$

$$\alpha_p = (1/2 + \alpha) \tag{209}$$

implying that Eq. (206) can be simplified and integrated into

$$p(\hat{y}|X, y) = \int p(\hat{y}, \sigma^2|X, y)\, d\sigma^2 = c\frac{\Gamma(\alpha_p)}{\beta_p^{\alpha_p}}\underbrace{\int \frac{\beta_p^{\alpha_p}}{\Gamma(\alpha_p)}(\sigma^2)^{-\alpha_p-1}e^{-\frac{\beta_p}{2\sigma^2}}\, d\sigma^2}_{==1} = c\frac{\Gamma(\alpha_p)}{\beta_p^{\alpha_p}}. \tag{210}$$

This gives us

$$p(\hat{y}|X, y) = (2\pi)^{-1/2}|\Sigma|^{-d/2}|\Lambda_p^{-1}|^{d/2}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha_p)}{\beta_p^{\alpha_p}}. \tag{211}$$