Course name: Math for ML
Name: Nguyen Thu Thao

Assignment 2

September 22, 2021

Question 1. *Proof: Multivariate Gaussian Distribution is normalized.*

My Answer 1.

Lets first introduce the Gaussian distribution over a variable $\mathbf{x} \in \mathbb{R}^D$,

$$\mathcal{N}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}$$

where μ is the mean and Σ is the covariance matrix. The mean takes the same dimensionality as the $\mathbf{x}$ and $\Sigma \in \mathbb{R}^D$. The characteristics of the Gaussian comes from the expression in the exponential,

$$\triangle^2 = (\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)$$

The matrix Σ can be taken to be symmetric. Now consider the eigenvector equation for the covariance matrix

$$\Sigma\mathbf{u}_i = \lambda_i\mathbf{u}_i$$

where $i = 1, \ldots, D$. Because   is a real, symmetric matrix its eigenvalues will be real, and its eigenvectors can be chosen to form an orthonormal set, so that

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij}$$

where I_{ij} is the i,j element of the identity matrix and satisfies

$$I_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$$

The covariance matrix Σ can be expressed as an expansion in terms of its eigenvectors in the form

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

and similarly the inverse covariance matrix Σ^{-1} can be expressed as

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

So that

$$\begin{aligned}\triangle^2 = (x-\mu)^T\Sigma^{-1}(x-\mu) &= \sum_{i=1}^{D} \frac{1}{\lambda_i}(x-\mu)^T u_i u_i^T (x-\mu) \\ &= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}, \text{ with } y_i = u_i^T(x-\mu)\end{aligned}$$

$$|\Sigma|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}$$

$$p(y) = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)}^{1/2} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} \Rightarrow \int_{-\infty}^{\infty} p(y)dy = \prod_{j=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_j)}^{1/2} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} dy_j = 1$$

Question 2. *Calculate the Marginal Gaussian Distribution*

My Answer 2.

In many situation, it will be convenient to work with the inverse of the covariance matrix , which is known as the precision matrix.

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & \dots & \Lambda_{1D} \\ \Lambda_{21} & \Lambda_{22} & \dots & \Lambda_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{D1} & \Lambda_{D2} & \dots & \Lambda_{DD} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1D} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{D1} & \Sigma_{D2} & \dots & \Sigma_{DD} \end{bmatrix}^{-1}$$

Write two-dimensional Gaussian using a precision matrix rather than a co-variance,

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}^{-1}\right)$$

Integrate out x_2 from the above and reach the marginal over x_1 as,

$$p(x_1) = \int p(x_1, x_2)\, \mathrm{d}x_2$$

The first thing we will do is to expand the exponent of the joint distribution,

$$\begin{aligned} E = &-\frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}} \Lambda_{11}(x_1 - \mu_1) - \frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}} \Lambda_{12}(x_2 - \mu_2) \\ &-\frac{1}{2}(x_2 - \mu_2)^{\mathrm{T}} \Lambda_{21}(x_1 - \mu_1) - \frac{1}{2}(x_2 - \mu_2)^{\mathrm{T}} \Lambda_{22}(x_2 - \mu_2) \end{aligned}$$

$$\begin{aligned} E = &-\frac{1}{2}\left(\left(x_2^{\mathrm{T}} \Lambda_{22} x_2 - 2x_2^{\mathrm{T}} \Lambda_{22}\left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right)\right. \\ &- 2x_1^{\mathrm{T}}\Lambda_{12}\mu_2 + 2\mu_1^{\mathrm{T}}\Lambda_{12}\mu_2 + \mu_2^{\mathrm{T}}\Lambda_{22}\mu_2 + x_1^{\mathrm{T}}\Lambda_{11}x_1 \\ &\left.-2x_1^{\mathrm{T}}\Lambda_{11}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{11}\mu_1\right) \\ = &-\frac{1}{2}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right)^{\mathrm{T}} \Lambda_{22}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right) \\ &+ \underbrace{\frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}x_1 - 2x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1\right)}_{A} \\ &- \underbrace{\frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{11}x_1 - 2x_1^{\mathrm{T}}\Lambda_{11}\mu_1 + \mu_1\Lambda_{11}\mu_1\right)}_{B} \end{aligned}$$

where we have used the fact that the co-variance matrix is symmetric such that $\Lambda_{12} = \Lambda_{21}^{\mathrm{T}}$ which allows us to,

$$x_1^{\mathrm{T}}\Lambda_{12}\mu_2 = x_1^{\mathrm{T}}\Lambda_{21}^{\mathrm{T}}\mu_2 = (\Lambda_{21}x_1)^{\mathrm{T}}\mu_2 = \mu_2^{\mathrm{T}}\Lambda_{21}x_1$$

We have:

$$\begin{aligned} A &= \frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}x_1 - 2x_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1 + \mu_1^{\mathrm{T}}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mu_1\right) \\ &= \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}\left(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right)(x_1 - \mu_1)\right) \\ B &= \frac{1}{2}\left(x_1^{\mathrm{T}}\Lambda_{11}x_1 - 2x_1^{\mathrm{T}}\Lambda_{11}\mu_1 + \mu_1\Lambda_{11}\mu_1\right) = \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}\Lambda_{11}(x_1 - \mu_1)\right) \\ A - B &= \frac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}\left(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} - \Lambda_{11}\right)(x_1 - \mu_1)\right) \end{aligned}$$

We have re-written the exponent as two separate terms, one term including x_2 and one which only includes x_1,

$$p(x_1, x_2) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{E_1} e^{E_2}$$

where,

$$\begin{aligned} E_1 &= -\tfrac{1}{2}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right)^{\mathrm{T}}\Lambda_{22}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right) \\ E_2 &= -\tfrac{1}{2}\left((x_1 - \mu_1)^{\mathrm{T}}\left(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right)(x_1 - \mu_1)\right. \end{aligned}$$

The formulation of the marginalisation we want to do we can exploit this structure.

$$p(x_1) = \int p(x_1, x_2)\,\mathrm{d}x_2 = \int \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{E_1} e^{E_2}\ \mathrm{d}x_2$$
$$= \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{E_2} \int e^{E_1}\ \mathrm{d}x_2$$

The way we have re-written the exponent only E_2 depends on x_2 while E_1 depends only on x_1

We will now proceed to integrate out x_2 from the first term in the exponent. Rather than doing this brute-force we can actually be a bit clever. If we look at E_1 we can see that it is also a quadratic form over x_2 just as the normal Gaussian,

$$E_1 = -\frac{1}{2}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right)^{\mathrm{T}} \Lambda_{22}\left(x_2 - \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)\right)$$

where we can think of $\left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)$ as the mean and Λ_{22} as the precision matrix.

$$\int \frac{1}{(2\pi)^{\frac{D_2}{2}} |\Lambda_{22}^{-1}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_2 - \tilde{\mu}_2)^{\mathrm{T}} \Lambda_{22}(x_2 - \tilde{\mu}_2)} \mathrm{d}x_2 = 1$$
$$\int e^{-\frac{1}{2}(x_2 - \tilde{\mu}_2)^{\mathrm{T}} \Lambda_{22}(x_2 - \tilde{\mu}_2)} \mathrm{d}x_2 = (2\pi)^{\frac{D_2}{2}} \left|\Lambda_{22}^{-1}\right|^{\frac{1}{2}}$$

where $\tilde{\mu_2} = \left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(x_1 - \mu_1)\right)$ and D_2 is the dimensionality of x_2. This means that we can re-write the expression as,

$$p(x_1) = (2\pi)^{\frac{D_2}{2}} \left|\Lambda_{22}^{-1}\right|^{\frac{1}{2}} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{E_2}$$
$$= \frac{1}{(2\pi)^{\frac{D-D2}{2}} \left|\Lambda_{22}^{-1}\right| - \frac{1}{2} |\Sigma|^{\frac{1}{2}}} e^{E_2}$$

Schur complement:

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A| \left|D - CA^{-1}B\right|$$
$$\Rightarrow |\Sigma| = |\Sigma_{11}| \left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\right| \Sigma_{12}|$$

The Schur complement of Λ_{22} is,

$$\Lambda_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

This means that we can simplify the terms that involves the derminants as follows,

$$\left|\Lambda_{22}^{-1}\right|^{-\frac{1}{2}} |\Sigma|^{\frac{1}{2}} = \left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|^{-\frac{1}{2}} |\Sigma_{11}|^{\frac{1}{2}} \left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|^{\frac{1}{2}}$$
$$= |\Sigma_{11}|^{\frac{1}{2}}$$

Now we can write down the full expression of the marginal distribution as follows,

$$p(x_1) = \frac{1}{(2\pi)^{\frac{D_1}{2}} |\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}}(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(x_1 - \mu_1)}$$

where we have used the fact that $D = D_1 + D_2$. The final step is now to re-write the expression in precision matrices in terms of a co-variance matrix. Again we will use the Schur complement to do this,

$$\Sigma_{11}^{-1} = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}$$

which leads to the final expression,

$$p(x_1) = \frac{1}{(2\pi)^{\frac{D_1}{2}} |\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_1 - \mu_1)^{\mathrm{T}} \Sigma_{11}^{-1}(x_1 - \mu_1)}$$

So, we have the formula Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}\left(\mathbf{x}_a \mid \mu_a, \Sigma_{aa}\right)$$

Question 3. *Calculate Conditional Gaussian distribution*

My Answer 3.

Conditional Distribution:

$$p\left(x_1, x_2\right) = p\left(x_1 \mid x_2\right) p\left(x_2\right)$$

The joint distribution,

$$\begin{aligned} p\left(x_1, x_2\right) &= \mathcal{N}\left(\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \\ &\propto e^{-\frac{1}{2}} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \end{aligned}$$

The marginal:

$$\begin{aligned} p\left(x_2\right) &= \frac{1}{(2\pi)^{\frac{D_2}{2}} \left|\Sigma_{22}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(x_2-\mu_2\right)^{\mathrm{T}} \Sigma_{22}^{-1}\left(x_2-\mu_2\right)} \\ &\propto e^{-\frac{1}{2}\left(x_2-\mu_2\right)^{\mathrm{T}} \Sigma_{22}^{-1}\left(x_2-\mu_2\right)} \end{aligned}$$

We will now use Schur complements to achieve this by expression the inverse of the full co-variance matrix decomposed in such a way that Σ_{22}^{-1} gets isolated. First lets look at the exponent of the joint distribution,

$$\begin{aligned} E &= -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ &= -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} \left(\Sigma/\Sigma_{22}\right)^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ &= -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \left(\Sigma/\Sigma_{22}\right)^{-1} & -\left(\Sigma/\Sigma_{22}\right)^{-1} \Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{21}\Sigma_{22}^{-1}\left(\Sigma/\Sigma_{22}\right)^{-1} & \Sigma_{22}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ &= -\frac{1}{2}\left(x - \left(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right)\right)^{\mathrm{T}} \left(\Sigma/\Sigma_{22}\right)^{-1} \left(x - \left(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right)\right) \\ &\underbrace{-\frac{1}{2}\left(x_2 - \mu_2\right)^{\mathrm{T}} \Sigma_{22}^{-1}\left(x_2 - \mu_2\right)}_{E_2} \end{aligned}$$

The last term in the expression above E_2 is exactly the exponent of the marginal distribution of x_2. Due to the product rule this means that we now know that the remaining term needs to be the exponent of the conditional Gaussian distribution. Therefore we only need to identify the parameters of a Gaussian to write down the posterior,

$$p\left(x_1 \mid x_2\right) \propto e_{\text{mean}}^{-\frac{1}{2}\left(x - \underbrace{\left(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}\right.}\left(x_2-\mu_2\right)\right)}\Bigg)^{\mathrm{T}} \Big(\underbrace{\Sigma/\Sigma_{22}}_{\text{covariance}} \Big)^{-1} \left(x - \left(\mu_1 + \Sigma_{21}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right)\right)\right)$$

from which we get the conditional distribution as,

$$p\left(x_1 \mid x_2\right) = \mathcal{N}\left(x_1 \mid \mu_1 + \Sigma_{21}\Sigma_{22}^{-1}\left(x_2 - \mu_2\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

where we have written out the Schur complement.