CS329 Homework #2

Course: Machine Learning(H)(CS329) - Instructor: Qi Hao

Name: Jianan Xie(谢嘉楠)

SID: 12110714

Question 1

(a) **[True or False]** If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian

Ans: True

(b) Consider a partitioning of the components of x into three groups x_a, x_b, and x_c, with a corresponding partitioning of the mean vector μ and of the covariance matrix Σ in the form

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}.$$

Find an expression for the conditional distribution $p(x_a|x_b)$ in which x_c has been marginalized out.

Ans:

For a joint Gaussian distribution $p(\mathbf{x}) = N(\mathbf{x}|\mu, \mathbf{\Sigma})$, where $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, $\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} = \Sigma^{-1}$,we have known that the conditional distribution $p(\mathbf{x_a}|\mathbf{x_b}) = N(\mathbf{x}|\mu_{\mathbf{a|b}}, \mathbf{\Lambda_{aa}^{-1}})$, where $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$. And the marginal distribution $p(\mathbf{x_a}) = N(\mathbf{x}|\mu_{\mathbf{a}}, \mathbf{\Sigma_{aa}})$.

So, first we consider $\mathbf{x} = \begin{pmatrix} \mathbf{x}_{a,b} \\ \mathbf{x}_c \end{pmatrix}$, where $\mathbf{x_{a,b}} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$. From above, we get the marginal distribution of $\mathbf{x_{a,b}}$ to make $\mathbf{x_c}$ marginalized out, then we get $p(\mathbf{x_{a,b}}) = N(\mathbf{x}|\mu_{\mathbf{a,b}}, \Sigma_{a,b})$,$\mu_{a,b} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$, $\Sigma_{a,b} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$. Next we find the conditional distribution $p(\mathbf{x_a}|\mathbf{x_b}) = N(\mathbf{x}|\mu_{\mathbf{a|b}}, \mathbf{\Lambda_{aa}^{-1}})$, where $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$, using what we learned above.

Question 2

Consider a joint distribution over the variable

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

whose mean and covariance are given by

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mu \\ \mathbf{A}\mu + \mathbf{b} \end{pmatrix}, \quad \mathrm{cov}[\mathbf{z}] = \begin{pmatrix} \mathbf{\Lambda^{-1}} & \mathbf{\Lambda^{-1}A}^{\mathrm{T}} \\ \mathbf{A\Lambda^{-1}} & \mathbf{L^{-1}} + \mathbf{A\Lambda^{-1}A}^{\mathrm{T}} \end{pmatrix}.$$

(a) Show that the marginal distribution $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda^{-1}})$.

Ans:

From what we have learned, when given a joint Gaussian distribution $p(\mathbf{x}) = N(\mathbf{x}|\mu, \mathbf{\Sigma})$, where $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, then the marginal distribution $p(\mathbf{x_a}) = N(\mathbf{x}|\mu_{\mathbf{a}}, \mathbf{\Sigma_{aa}})$. Here, $\mu_x = \mu$ and $\Sigma_{xx} = \Lambda^{-1}$, Therefore, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda^{-1}})$.

(b) Show that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is given by $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L^{-1}})$.

Ans:

From what we have learned, when given a joint Gaussian distribution $p(\mathbf{x}) = N(\mathbf{x}|\mu, \mathbf{\Sigma})$, where $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, then the conditional distribution $p(\mathbf{x_a}|\mathbf{x_b}) = N(\mathbf{x}|\mu_{\mathbf{a|b}}, \mathbf{\Lambda_{aa}^{-1}})$, where $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$. Therefore, $\Sigma_{y|x} = \Lambda_{yy}^{-1} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} = (\mathbf{L^{-1}} + \mathbf{A\Lambda^{-1}A}^{\mathrm{T}}) - (\mathbf{A\Lambda^{-1}})(\mathbf{\Lambda^{-1}})^{-1}(\mathbf{\Lambda^{-1}A}^{\mathrm{T}}) = \mathbf{L^{-1}}$ and $\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu) = (\mathbf{A}\mu + \mathbf{b}) + (\mathbf{A\Lambda^{-1}})(\mathbf{\Lambda^{-1}})^{-1}(x - u) = \mathbf{Ax} + \mathbf{b}$. So, $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L^{-1}})$.

Question 3

Show that the covariance matrix Σ that maximizes the log likelihood function is given by the sample covariance

$\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu)$.

Is the final result symmetric and positive definite (provided the sample covariance is nonsingular)?

> **Hints**
>
> *(a)* To find the maximum likelihood solution for the covariance matrix of a multivariate Gaussian, we need to maximize the log likelihood function with respect to Σ. The log likelihood function is given by
>
> $\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu)$.
>
> *(b)* The derivative of the inverse of a matrix can be expressed as
>
> $\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$
>
> We have the following properties
>
> $\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{A}) = \mathbf{I}, \quad \frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A^{-1}})^{\mathrm{T}}$.

Ans:

From the hint(a), we only need to maximize the log likelihood function with respect to Σ. Here, we will use two useful identities for computing gradients. First is $\frac{\partial \mathbf{a}^{\mathrm{T}}\mathbf{X^{-1}b}}{\partial \mathbf{X}} = -(\mathbf{X}^{-1})^{\mathrm{T}}\mathbf{ab}^{\mathrm{T}}(\mathbf{X}^{-1})^{\mathrm{T}}$, and the second is $\frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A^{-1}})^{\mathrm{T}}$ in hint(b).

So, we derive that

$$\begin{aligned}
\frac{\partial \ln p(\mathbf{X}|\mu, \Sigma)}{\partial \Sigma} &= -\frac{N}{2}\frac{\partial}{\partial \Sigma}\ln|\Sigma| - \frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu) \\
&= -\frac{N}{2}(\Sigma^{-1})^{\mathrm{T}} + \frac{1}{2}\sum_{n=1}^{N}(\Sigma^{-1})^{\mathrm{T}}(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}}(\Sigma^{-1})^{\mathrm{T}} \\
&= -\frac{N}{2}(\Sigma^{-1}) + \frac{1}{2}\sum_{n=1}^{N}(\Sigma^{-1})(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}}(\Sigma^{-1}) \qquad (\text{Since } \Sigma \text{ is symmetric}) \\
&= 0
\end{aligned}$$

Then we get $-N\Sigma + \sum_{n=1}^{N}(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}} = 0$ by multiplying Σ on left side and on right side in sequence. Hence, we derive the $\Sigma_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}}$ which is the sample covariance matirx. Since the sample covariance is nonsingular, the final result is symmetric and positive definite.

Question 4

(a) Derive an expression for the sequential estimation of the variance of a univariate Gaussian distribution, by starting with the maximum likelihood expression

$\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$.

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

Ans:

we will denote $\sigma_{ML}^{(N)}$ as maximum likelihood estimator of σ when it is based on N observations

$$\begin{aligned}\sigma_{\mathrm{ML}}^{2(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2. \\ &= \frac{1}{N}(x_N-\mu)^2 + \frac{1}{N}\sum_{n=1}^{N-1}(x_n-\mu)^2 \\ &= \frac{1}{N}(x_N-\mu)^2 + \frac{N-1}{N}\sigma_{\mathrm{ML}}^{2(N-1)} \\ &= \sigma_{\mathrm{ML}}^{2(N-1)} + \frac{1}{N}((x_N-\mu)^2 - \sigma_{\mathrm{ML}}^{2(N-1)})\end{aligned}$$

by Robbins-Monro sequential estimation formula,

$$\theta^{(N)} = \theta^{(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}[-\ln p(x_N|\theta^{(N-1)})]$$

substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives,

$$\begin{aligned}\sigma_{ML}^{2(N)} &= \sigma_{ML}^{2(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\sigma_{ML}^{2(N-1)}}[-\ln p(x_N|\sigma_{ML}^{2(N-1)})] \\ &= \sigma_{ML}^{2(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\sigma_{ML}^{2(N-1)}}[\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\sigma_{ML}^{2(N-1)} + \frac{(x_N-\mu)^2}{2\sigma_{ML}^{2(N-1)}}] \\ &= \sigma_{ML}^{2(N-1)} - \alpha_{N-1}(\frac{1}{2\sigma_{ML}^{2(N-1)}} - \frac{(x_N-\mu)^2}{2\sigma_{ML}^{4(N-1)}}) \\ &= \sigma_{ML}^{2(N-1)} + \frac{\alpha_{N-1}}{2\sigma_{ML}^{4(N-1)}}((x_N-\mu)^2 - \sigma_{ML}^{2(N-1)})\end{aligned}$$

Thus, we take $\alpha_N = \frac{2\sigma_{ML}^{4(N)}}{N+1}$ to make the result of Robbins-Monro same with $\sigma_{\mathrm{ML}}^{2(N-1)} + \frac{1}{N}((x_N-\mu)^2 - \sigma_{\mathrm{ML}}^{2(N-1)})$.

(b) Derive an expression for the sequential estimation of the covariance of a multivariate Gaussian distribution, by starting with the maximum likelihood expression

$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\mu_{\mathrm{ML}})(\mathbf{x}_n-\mu_{\mathrm{ML}})^{\mathrm{T}}$.

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

> ### Hints
>
> *(a)* Consider the result $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ for the maximum likelihood estimator of the mean μ_{ML}, which we will denote by $\mu_{\mathrm{ML}}^{(N)}$ when it is based on N observations. If we dissect out the contribution from the final data point $\mathbf{x}_N$, we obtain
>
> $\mu_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{\mathrm{ML}}^{(N-1)}$
>
> *(b)* Robbins-Monro for maximum likelihood
>
> $\theta^{(N)} = \theta^{(N-1)} + a_{(N-1)}\frac{\partial}{\partial\theta^{(N-1)}}\ln p(x_N|\theta^{(N-1)})$.

Ans:

we will denote $\Sigma_{ML}^{(N)}$ as maximum likelihood estimator of Σ when it is based on N observations

$$\begin{aligned}\Sigma_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\mu_{\mathrm{ML}})(\mathbf{x}_n-\mu_{\mathrm{ML}})^{\mathrm{T}} \\ &= \frac{1}{N}(\mathbf{x}_N-\mu_{\mathrm{ML}})(\mathbf{x}_N-\mu_{\mathrm{ML}})^{\mathrm{T}} + \frac{1}{N}\sum_{n=1}^{N-1}(\mathbf{x}_n-\mu_{\mathrm{ML}})(\mathbf{x}_n-\mu_{\mathrm{ML}})^{\mathrm{T}} \\ &= \frac{1}{N}(\mathbf{x}_N-\mu_{\mathrm{ML}})(\mathbf{x}_N-\mu_{\mathrm{ML}})^{\mathrm{T}} + \frac{N-1}{N}\Sigma_{ML}^{(N-1)} \\ &= \Sigma_{ML}^{(N-1)} + \frac{1}{N}((\mathbf{x}_N-\mu_{\mathrm{ML}})(\mathbf{x}_N-\mu_{\mathrm{ML}})^{\mathrm{T}} - \Sigma_{ML}^{(N-1)})\end{aligned}$$

by Robbins-Monro sequential estimation formula,

$$\theta^{(N)} = \theta^{(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}[-\ln p(x_N|\theta^{(N-1)})]$$

substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives,

$$\begin{aligned}
\Sigma_{\mathrm{ML}}^{(N)} &= \Sigma_{\mathrm{ML}}^{(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\Sigma_{\mathrm{ML}}^{(N-1)}}[-\ln p(\mathbf{x}_N|\Sigma_{\mathrm{ML}}^{(N-1)})] \\
&= \Sigma_{\mathrm{ML}}^{(N-1)} - \alpha_{N-1}\frac{\partial}{\partial\Sigma_{\mathrm{ML}}^{(N-1)}}[\frac{ND}{2}\ln(2\pi) + \frac{N}{2}\ln|\Sigma_{\mathrm{ML}}^{(N-1)}| + \frac{1}{2}(\mathbf{x}_N - \mu_{\mathrm{ML}})^{\mathrm{T}}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}(\mathbf{x}_N - \mu_{\mathrm{ML}})] \\
&= \Sigma_{\mathrm{ML}}^{(N-1)} - \alpha_{N-1}[\frac{N}{2}((\Sigma_{\mathrm{ML}}^{(N-1)})^{-1})^{\mathrm{T}} - \frac{1}{2}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}(\mathbf{x}_N - \mu_{\mathrm{ML}})(\mathbf{x}_N - \mu_{\mathrm{ML}})^{\mathrm{T}}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}] \\
&= \Sigma_{\mathrm{ML}}^{(N-1)} - \alpha_{N-1}[\frac{N}{2}((\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}) - \frac{1}{2}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}(\mathbf{x}_N - \mu_{\mathrm{ML}})(\mathbf{x}_N - \mu_{\mathrm{ML}})^{\mathrm{T}}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}] \\
&= \Sigma_{\mathrm{ML}}^{(N-1)} + \alpha_{N-1}\frac{N}{2}(\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}[\frac{1}{N}((\mathbf{x}_N - \mu_{\mathrm{ML}})(\mathbf{x}_N - \mu_{\mathrm{ML}})^{\mathrm{T}} - \Sigma_{ML}^{(N-1)})](\Sigma_{\mathrm{ML}}^{(N-1)})^{-1}
\end{aligned}$$

Thus, we take $\alpha_N = \frac{2\Sigma_{ML}^{2(N)}}{N+1}$ to make the result of Robbins-Monro same with $\Sigma_{ML}^{(N-1)} + \frac{1}{N}((\mathbf{x}_N - \mu_{\mathrm{ML}})(\mathbf{x}_N - \mu_{\mathrm{ML}})^{\mathrm{T}} - \Sigma_{ML}^{(N-1)})$.

Question 5

Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $N(x|\mu, \Sigma)$ in which the covariance Σ is known and for which we wish to infer the mean μ from a set of observations $\mathbf{X} = \{x_1, x_2, \ldots\ldots, x_N\}$. Given a prior distribution $p(\mu) = N(\mu|\mu_0, \Sigma_0)$, find the corresponding posterior distribution $p(\mu|\mathbf{X})$.

Ans:

The likelihood function is $p(\mathbf{X}|\mu) = \prod_{n=1}^{N} p(\mathbf{x_n}|\mu) = \frac{1}{(2\pi)^{DN/2}|\Sigma|^{N/2}} exp\{\sum_{n=1}^{N} -\frac{1}{2}(\mathbf{x_n} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x_n} - \mu)\}$, and the prior distribution is $p(\mu) = N(\mu|\mu_0, \Sigma_0)$. Here, the prior *p*(*μ*) is given by a Gaussian, it will be a conjugate distribution for this likelihood function because the corresponding posterior will be a product of two exponentials of quadratic functions of μ and hence will also be Gaussian. Thus, we suppose $p(\mu|\mathbf{X}) = N(\mu|\mu_N, \Sigma_N)$. According to Bayesian Inference $p(\mu|\mathbf{X}) \propto p(\mathbf{X}|\mu)p(\mu)$, we focus on the exponential term and rearrange it according to μ.

We have the exponential term of $p(\mathbf{X}|\mu)p(\mu)$ like this:

$$\begin{aligned}
&\sum_{n=1}^{N} -\frac{1}{2}(\mathbf{x_n} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x_n} - \mu) - \frac{1}{2}(\mu - \mu_{\mathbf{0}})^{\mathrm{T}}\Sigma_0^{-1}(\mu - \mu_{\mathbf{0}}) \\
&= -\frac{1}{2}\mu^{\mathrm{T}}(N\Sigma^{-1} + \Sigma_0^{-1})\mu + \frac{1}{2}\mu^{\mathrm{T}}(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x_n} + \Sigma_0^{-1}\mu_0) + \frac{1}{2}(\sum_{n=1}^{N}\mathbf{x_n}^{\mathrm{T}}\Sigma^{-1} + \mu_0^{\mathrm{T}}\Sigma^{-1})\mu - \frac{1}{2}(\sum_{n=1}^{N}\mathbf{x_n}^{\mathrm{T}}\mathbf{x_n} + \mu_0^{\mathrm{T}}\mu_0)
\end{aligned}$$

and exponential term of $p(\mu|\mathbf{X})$ like this:

$$\begin{aligned}
&-\frac{1}{2}(\mu - \mu_N)^{\mathrm{T}}\Sigma_N^{-1}(\mu - \mu_N) \\
&= -\frac{1}{2}\mu^{\mathrm{T}}\Sigma_N^{-1}\mu + \frac{1}{2}\mu^{\mathrm{T}}\Sigma_N^{-1}\mu_N + \frac{1}{2}\mu_N^{\mathrm{T}}\Sigma_N^{-1}\mu - \frac{1}{2}\mu_N^{\mathrm{T}}\mu_N
\end{aligned}$$

So, we have

$$\begin{aligned}
\Sigma_N^{-1} &= N\Sigma^{-1} + \Sigma_0^{-1} \\
\mu_N &= \Sigma_N(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x_n} + \Sigma_0^{-1}\mu_0) = (N\Sigma^{-1} + \Sigma_0^{-1})^{-1}(\Sigma^{-1}\sum_{n=1}^{N}\mathbf{x_n} + \Sigma_0^{-1}\mu_0)
\end{aligned}$$