A Review on Multivariate Gaussian Distribution And An Optimization Point of View for The Conditional Distribution

J.R

August 20, 2021

In this note, I want to make a review on multivariate Gaussian Distribution, including the conditional distribution, marginal distribution, covariance matrix and information matrix of multivariate Gaussian distribution. There is nothing fancy in this note and some proof in this note is not the simplest way to make proofs, it basicly list some important fact of Gaussian distribution and the proof. However, I hope it can provide you a new perspective to think of Gaussian distribution. Hopefully, this can be a good complementary material for Chapter 7 of the book *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller.

Contents

1 Univariate Gaussian Distribution

The Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is defined by

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}) \tag{1.1}$$

I assume you are familiar with these bell-shaped curves, but for the sake of completeness, let me show you those curves with $\mu = 0$.

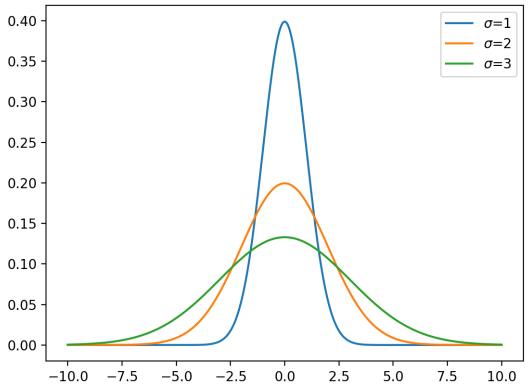


Figure 1: 1d Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$

The first thing I want to show is how to integrate this function. There's no close-form solution for the CDF of a Gaussian distribution, but we can still integrate over the whole real line. The trick is, the same as what we used in the note of Conjugate Gradient method, change of coordinate. Let's integrate a standard normal $\mathcal{N}(0, 1)$.

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \tag{1.2}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dy dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dy dx \tag{1.3}$$

Now make a change of coordinate from cartesian coordinate to polar coordinate, then (1.3) becomes

$$\int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}r^2} r dr d\theta \qquad \text{let x} = \frac{r^2}{2} \tag{1.4}$$

$$= \int_0^{2\pi} \int_0^{\infty} e^{-x} dx dr$$

$$= 2\pi$$

Thus $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = 2\pi$ and the standard normal integrates to 1.

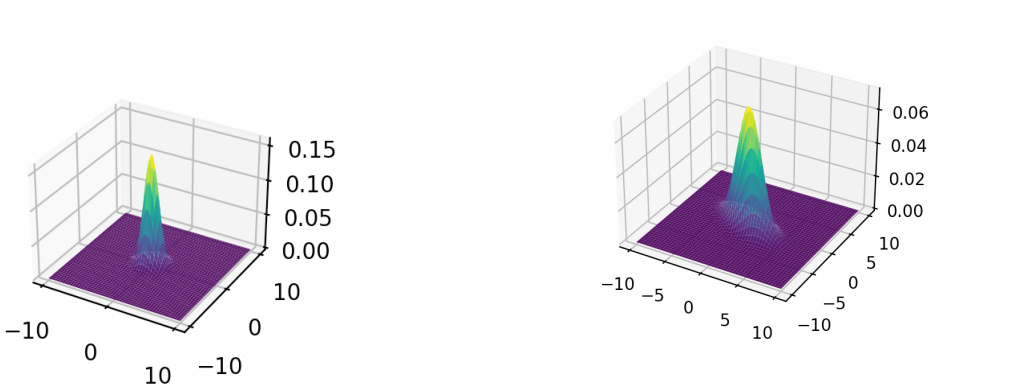
2 Multivariate Gaussian Distribution

The multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is defined as

$$p_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{2.1}$$

where μ is the expectation, Σ is a positive definite covariance matrix and n is the dimension of random variable X.

2d Gaussian distributions with $\mu = 0$ and different Σ are shown in figure 2



(a) $\Sigma = I$

(b) $\Sigma = \begin{pmatrix} 1/\sqrt{5} & -2\sqrt{5} \\ 2\sqrt{5} & 1\sqrt{5} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} 1\sqrt{5} & -2\sqrt{5} \\ 2\sqrt{5} & 1\sqrt{5} \end{pmatrix}^T$

Figure 2: 2D Gaussian distributions with $\mu = 0$ and different Σ

In the case of $n = 2$, when the covariance matrix is the identity matrix, it's obvious the contour lines of $p_X(x)$ are a group of concentric circles, and when we integrate over the whole plane $\mathbb{R}^2$, we basicly do the same thing as we did in (1.2).

When Σ is not identity, the quadratic form in the exponential $(x-\mu)^T\Sigma^{-1}(x-\mu)$ indicates a ellipsoid. So the contour lines of Gaussian are a group of concentric ellipsoids centered at μ. We can make a change of variable to the quadratic form $v = \Sigma^{-1/2}(x-\mu)$, $x - \mu = \Sigma^{1/2}v$. The exponential term becomes $\exp(-\frac{1}{2}v^Tv)$, which has the form of a standard normal which integrates to 2π, as shown in (1.2). However, the term $\frac{1}{(2\pi)^{n/2}}$ itself does not give use the normalizing constant. When we apply a linear transformation $x - \mu = \Sigma^{1/2}v$, we have also scaled the area over which we integrate the pdf. Each unit area ds is scaled by $|\Sigma^{1/2}|$, thus we want to scale each unit probability $P(x \in ds')$ down by $|\Sigma^{1/2}| = |\Sigma|^{1/2}$ so that the pdf integrate to 1. The scaling is shown in figure 3.

3 Conditional of Multivariate Gaussian

Conditional distribution of a multivariate Gaussian comes up from time to time in all kinds of contexts. For example, linear Gaussian Bayesian network and factor analysis. However, the formulas for conditional expectation and conditional covariance matrix is quite mysterious. In this section, I want to derive the conditional expectation with a convex optimization point of view and derive the conditional covariance by simply inverting a block matrix. This may not be the easiest way to derive the formulas, but I hope these can provide you a new perspective.

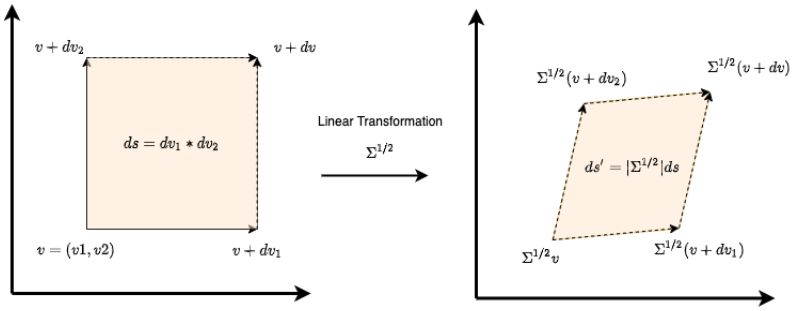


Figure 3

We assume the random variable $X = [X_1, X_2]$ has a joint distribution $X = [X_1, X_2] \sim \mathcal{N}(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{3.1}$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{3.2}$$

and $\Sigma_{12} = \Sigma_{21}^T$.

First, we can easily conclude that the conditional distribution of a multivariate Gaussian distribution is also a Gaussian distribution because when we fix X_2, the terms not relevant to X_1 comes out of the exponential as a constant, and the remaining terms in the exponential is still a quadratic form. So it still have a Gaussian pattern, except that we need a different normalizing constant.

3.1 Conditional Expectation of Multivariate Gaussian

Once we have the conclusion that the conditional distribution of a multivariate Gaussian distribution is also a Gaussian distribution, we can formulate the problem of finding the conditional expectation as a optimization problem since a Gaussian PDF achieve the maximum at the expectation. Thus, when we condition on $X_2 = x_2$, we can write down the optimization problem

$$\min_{X_1} . \frac{1}{2} \begin{bmatrix} X_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} X_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \tag{3.3}$$

which is a convex problem since Σ is a positive definite matrix.

If we don't have the inverse if Σ, we can easily write the quadratic form with block notation of Σ and solve the problem by simply setting gradient to 0. However, with the inverse sign, we need to resort to more advanced optimization techniques. We can introduce a "redundant" equality constraint $X_2 = x_2$, and formulate the optimization problem as

$$\begin{aligned} \min_{X_1, X_2} . \quad & \frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu) \\ \text{s.t.} \quad & X_2 = x_2 \end{aligned} \tag{3.4}$$

where $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. The Lagrangian is

$$\mathcal{L}(X_1, X_2, v) = \frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) + v^T(X_2 - x_2) \tag{3.5}$$

$$\nabla_X \mathcal{L}(X_1, X_2, v) = \Sigma^{-1}(X-\mu) + \begin{bmatrix} 0 \\ v \end{bmatrix} := 0 \tag{3.6}$$

$$X^*(v) = -\Sigma \begin{bmatrix} 0 \\ v \end{bmatrix} + \mu \tag{3.7}$$

where $X_1^*(v) = -\Sigma_{12}v + \mu_1$, $X_2^*(v) = -\Sigma_{22}v + \mu_2$, thus the dual function is

$$\begin{aligned} g(v) &= \inf_{X_1, X_2} \mathcal{L}(X_1, X_2, v) \\ &= \frac{1}{2}\begin{bmatrix} 0 & v \end{bmatrix}\Sigma\Sigma^{-1}\Sigma\begin{bmatrix} 0 \\ v \end{bmatrix} + v^T(-\Sigma_{22}v + \mu_2 - x_2) \\ &= -\frac{1}{2}v^T\Sigma_{22}v + v^T(\mu_2 - x_2) \end{aligned} \tag{3.8}$$

The dual function, as convex Lagrange duality suggests, is a concave function. And the inverse of Σ no longer exists. We can easily compute the dual optimal by setting the gradient to 0,

$$\begin{aligned} \nabla_v g(v) &= -\Sigma_{22}v + \mu_2 - x_2 := 0 \\ v^* &= -\Sigma_{22}^{-1}(x_2 - \mu_2) \end{aligned} \tag{3.9}$$

Since there's no inequality constrains in problem(3.4), slater condition (trivially) holds, so $p^* = d^*$,

$$p^* = d^* = g(v^*) = \mathcal{L}(X_1(v^*), X_2(v^*), v^*)$$

$$X^* = -\Sigma \begin{bmatrix} 0 \\ v^* \end{bmatrix} + \mu$$

$$X_1^* = -\Sigma_{12}v^* + \mu_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{3.10}$$

$$X_2^* = -\Sigma_{22}v^* + \mu_2 = x_2 \tag{3.11}$$

Thus we have the conditional expectation of a multivariate Gaussian distribution.

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \tag{3.12}$$

(we don't need to compute (3.11), but it's good to double-check the feasibility to ensure that we did the correct calculation.)

3.2 Conditional Covariance Matrix of Multivariate Gaussian

Now we have the conditional expectation of a multivariate Gaussian, what we need now is the conditional covariance matrix. Since we know that the conditional distribution of a Gaussian is also a Gaussian, we can write the PDF of $X_1|X_2$

$$f_{X_1|X_2}(X_1|X_2) = C \cdot \exp(-\frac{1}{2}(X_1 - \mu_{1|2})^T\Sigma_{1|2}^{-1}(X_1 - \mu_{1|2})) \tag{3.13}$$

where C is the normalizing constant. We now have $\mu_{1|2}$ according to (3.12). We now want to find $\Sigma_{1|2}^{-1}$, the coefficient of the second-order term for X_2. From (3.3), we can see the second-order terms for X_2 only relates to the upper-left block of Σ^{-1}. Thus, $\Sigma_{1|2}^{-1}$ is the upper-left block of Σ^{-1}, in other words, we want to find $\Sigma_{1|2}$, which is the inverse of upper-left block of Σ^{-1}.

The problem is how do we symbolically compute Σ^{-1}. Actually, we can do this from a linear equation point of view.

First, we can write down a linear equation about Σ

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}u\\ v\end{bmatrix} \tag{3.14}$$

Since Σ is a positive definite matrix, if we can write down a matrix A such that $A\begin{bmatrix}u\\ v\end{bmatrix} = \begin{bmatrix}x\\ y\end{bmatrix}$, then A must be inverse of Σ. And we can come up with such a matrix A by block elimination.

First, we can we eliminate y by expressing y in terms of x. According to the second row block

$$\begin{aligned}\Sigma_{21}x + \Sigma_{22}y &= v\\ y &= \Sigma_{22}^{-1}(v - \Sigma_{21}x)\end{aligned} \tag{3.15}$$

and according to the first row, we have

$$\Sigma_{11}x + \Sigma_{12}y = u \tag{3.16}$$

plug (3.15) into (3.16)

$$\begin{aligned}\Sigma_{11}x + \Sigma_{12}\Sigma_{22}^{-1}(v - \Sigma_{21}x) &= u\\ x &= (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(u - \Sigma_{12}\Sigma_{22}^{-1}v)\end{aligned} \tag{3.17}$$

then plug (3.17) into (3.15), we have

$$\begin{aligned}y &= \Sigma_{22}^{-1}\Big(v - \Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(u - \Sigma_{12}\Sigma_{22}^{-1}v)\Big)\\ &= -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}u + \Big[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Big]v\end{aligned} \tag{3.18}$$

Before anything else, I want you to pause for 5 minutes and just appreciate the beauty of math. Look at the coefficient for v in (3.18), it show some "recursive-symmetric" pattern.

According, to (3.17) and (3.18), we can write down a linear equation

$$\begin{aligned}A\begin{bmatrix}u\\ v\end{bmatrix} &= \begin{bmatrix}x\\ y\end{bmatrix}\\ A &= \begin{bmatrix}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{bmatrix}\end{aligned} \tag{3.19}$$

thus A must be Σ_{-1}. So the quadratic terms of X_1 can be written as $X_1^T(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}X_1$ and the conditional covariance matrix is

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{3.20}$$

(3.20) is called a/the schur complement of Σ.

3.3 Exploiting the Structure Further

Now we have the conditional covariance matrix. However, (3.18) looks very cumbersome, can we simplify it a little bit? Actually yes. We get (3.18) by expressing y in terms of x, what if we express x in terms of y first? When we eliminate y, we get a linear equation like the following

$$B\begin{bmatrix}u\\v\end{bmatrix}=\begin{bmatrix}x\\y\end{bmatrix}$$
$$B=\begin{bmatrix}\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\\ -(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & (\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\end{bmatrix} \tag{3.21}$$

Let's compare (3.19) and (3.20), A and B are both inverse of Σ, they must be the same. Thus we can get two not obvious identity,

$$(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}=\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} \tag{3.22}$$
$$-(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}=-\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \tag{3.23}$$

Also, it seems to be natural that we expect similar formulas for $\Sigma_{2|1}$ and $\Sigma_{1|2}$.

4 Marginal of Multivariate Gaussian

Now we have the conditional expectation, conditional covariance matrix, some identities according to (3.22) and (3.23), and the block representation of the inverse of covariance matrix according to (3.19) and (3.21). We can use these to derive the marginal of a multivariate Gaussian. First, let's write the inverse of the covariance matrix Σ in blocks

$$\Sigma^{-1}=J=\begin{bmatrix}J_{11} & J_{12}\\ J_{21} & J_{22}\end{bmatrix} \tag{4.1}$$

The PDF of the multivariate Gaussian can be written as

$$\begin{aligned}f_X(x)&=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^TJ(x-\mu)\right]\\&=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}[(x_1-\mu_1)^TJ_{11}(x_1-\mu_1)+2(x_1-\mu_1)^TJ_{12}(x_2-\mu_2)+(x_2-\mu_2)^TJ_{22}(x_1-\mu_1)]\right]\end{aligned} \tag{4.2}$$

Let's only look at the quadratic term in $\exp(\cdot)$ and plug in the block representation of J's with what we got in (3.19) and (3.21).

$$\begin{aligned}&(x_1-\mu_1)^TJ_{11}(x_1-\mu_1)+2(x_1-\mu_1)^TJ_{12}(x_2-\mu_2)+(x_2-\mu_2)^TJ_{22}(x_1-\mu_1)\\=&(x_1-\mu_1)^T\left[\Sigma_{11}^{-1}+\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1}\right](x_1-\mu_1)\\&-2(x_1-\mu_1)^T\left[\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\right](x_2-\mu_2)\\&+(x_2-\mu_2)^T(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(x_1-\mu_1)\\=&(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\\&+(x_2-\mu_2-\Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1))^T(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}(x_2-\mu_2-\Sigma_{21}\Sigma_{11}^{-1}(x_1-\mu_1))\\=&(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)+(x_2-\mu_{2|1})^T\Sigma_{2|1}^{-1}(x_2-\mu_{2|1})\end{aligned} \tag{4.3}$$

Then (4.2) can be written as

$$\begin{aligned} f_X(x) =& \frac{1}{(2\pi)^{n_1/2}|\Sigma_{11}|^{1/2}} \frac{1}{(2\pi)^{n_2/2}|\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{1/2}} \\ & \exp\Big(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\Big) \\ & \exp\Big(-\frac{1}{2}(x_2-\mu_{2|1})^T\Sigma_{2|1}^{-1}(x_1-\mu_{2|1})\Big) \\ =& \frac{1}{(2\pi)^{n_1/2}|\Sigma_{11}|^{1/2}} \exp\Big(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\Big) f_{X_2|X_1}(x_2|x_1) \end{aligned} \tag{4.4}$$

where n_1 is the dimension of X_1 and n_2 is the dimension of X_2 (i.e. $n_1+n_2=n$). In the first equal sign we use the fact that the determinant of a psd matrix is equal to the determinant of the upper-left block times the determinant of the schur-complement. Using the law of total probability, we can easily conclude that

$$\begin{aligned} f_{X_1}(x_1) &= \frac{f_{X_1,X_2}(X_1,X_2)}{f_{X_2|X_1}(x_2|x_1)} \\ &= \frac{1}{(2\pi)^{n_1/2}|\Sigma_{11}|^{1/2}} \exp\Big(-\frac{1}{2}(x_1-\mu_1)^T\Sigma_{11}^{-1}(x_1-\mu_1)\Big) \end{aligned}$$

Thus we conclude that in a joint distribution

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\Big(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \Big)$$

The marginal distribution is

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$$