The Multivariate Gaussian Distribution

Chuong B. Do

October 10, 2008

A vector-valued random variable $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\mu \in \mathbf{R}^n$ and covariance matrix $\Sigma \in \mathbf{S}^n_{++}$[1] if its probability density function[2] is given by

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right).$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$. In these notes, we describe multivariate Gaussians and some of their basic properties.

1 Relationship to univariate Gaussians

Recall that the density function of a **univariate normal (or Gaussian) distribution** is given by

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right).$$

Here, the argument of the exponential function, $-\frac{1}{2\sigma^2}(x-\mu)^2$, is a quadratic function of the variable x. Furthermore, the parabola points downwards, as the coefficient of the quadratic term is negative. The coefficient in front, $\frac{1}{\sqrt{2\pi}\sigma}$, is a constant that does not depend on x; hence, we can think of it as simply a "normalization factor" used to ensure that

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right) = 1.$$

[1] Recall from the section notes on linear algebra that $\mathbf{S}^n_{++}$ is the space of symmetric positive definite $n \times n$ matrices, defined as

$$\mathbf{S}^n_{++} = \left\{ A \in \mathbf{R}^{n \times n} : A = A^T \text{ and } x^T A x > 0 \text{ for all } x \in \mathbf{R}^n \text{ such that } x \neq 0 \right\}.$$

[2] In these notes, we use the notation $p(\bullet)$ to denote density functions, instead of $f_X(\bullet)$ (as in the section notes on probability theory).

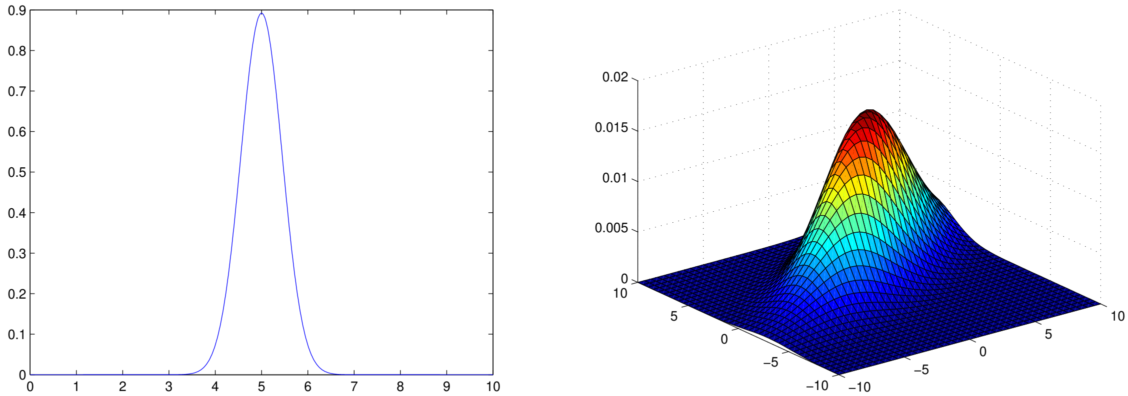

Figure 1: The figure on the left shows a univariate Gaussian density for a single variable X. The figure on the right shows a multivariate Gaussian density over two variables X_1 and X_2.

In the case of the multivariate Gaussian density, the argument of the exponential function, $-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$, is a **quadratic form** in the vector variable x. Since Σ is positive definite, and since the inverse of any positive definite matrix is also positive definite, then for any non-zero vector z, $z^T\Sigma^{-1}z > 0$. This implies that for any vector $x \neq \mu$,

$$(x-\mu)^T\Sigma^{-1}(x-\mu) > 0$$
$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) < 0.$$

Like in the univariate case, you can think of the argument of the exponential function as being a downward opening quadratic bowl. The coefficient in front (i.e., $\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}$) has an even more complicated form than in the univariate case. However, it still does not depend on x, and hence it is again simply a normalization factor used to ensure that

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx_1dx_2\cdots dx_n = 1.$$

2 The covariance matrix

The concept of the **covariance matrix** is vital to understanding multivariate Gaussian distributions. Recall that for a pair of random variables X and Y, their **covariance** is defined as

$$Cov[X,Y] = E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y].$$

When working with multiple variables, the covariance matrix provides a succinct way to summarize the covariances of all pairs of variables. In particular, the covariance matrix, which we usually denote as Σ, is the $n \times n$ matrix whose (i,j)th entry is $Cov[X_i, X_j]$.

The following proposition (whose proof is provided in the Appendix A.1) gives an alternative way to characterize the covariance matrix of a random vector X:

Proposition 1. *For any random vector X with mean μ and covariance matrix Σ,*

$$\Sigma = E[(X-\mu)(X-\mu)^T] = E[XX^T] - \mu\mu^T. \tag{1}$$

In the definition of multivariate Gaussians, we required that the covariance matrix Σ be symmetric positive definite (i.e., $\Sigma \in \mathbf{S}^n_{++}$). Why does this restriction exist? As seen in the following proposition, the covariance matrix of *any* random vector must always be symmetric positive semidefinite:

Proposition 2. *Suppose that Σ is the covariance matrix corresponding to some random vector X. Then Σ is symmetric positive semidefinite.*

Proof. The symmetry of Σ follows immediately from its definition. Next, for any vector $z \in \mathbf{R}^n$, observe that

$$\begin{aligned}
z^T\Sigma z &= \sum_{i=1}^n \sum_{j=1}^n (\Sigma_{ij} z_i z_j) && (2)\\
&= \sum_{i=1}^n \sum_{j=1}^n (Cov[X_i, X_j] \cdot z_i z_j)\\
&= \sum_{i=1}^n \sum_{j=1}^n (E[(X_i - E[X_i])(X_j - E[X_j])] \cdot z_i z_j)\\
&= E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - E[X_i])(X_j - E[X_j]) \cdot z_i z_j\right]. && (3)
\end{aligned}$$

Here, (2) follows from the formula for expanding a quadratic form (see section notes on linear algebra), and (3) follows by linearity of expectations (see probability notes).

To complete the proof, observe that the quantity inside the brackets is of the form $\sum_i \sum_j x_i x_j z_i z_j = (x^T z)^2 \geq 0$ (see problem set #1). Therefore, the quantity inside the expectation is always nonnegative, and hence the expectation itself must be nonnegative. We conclude that $z^T \Sigma z \geq 0$. □

From the above proposition it follows that Σ must be symmetric positive semidefinite in order for it to be a valid covariance matrix. However, in order for Σ^{-1} to exist (as required in the definition of the multivariate Gaussian density), then Σ must be invertible and hence full rank. Since any full rank symmetric positive semidefinite matrix is necessarily symmetric positive definite, it follows that Σ must be symmetric positive definite.

3 The diagonal covariance matrix case

To get an intuition for what a multivariate Gaussian is, consider the simple case where $n = 2$, and where the covariance matrix Σ is diagonal, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

In this case, the multivariate Gaussian density has the form,

$$\begin{aligned} p(x;\mu,\Sigma) &= \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \right) \\ &= \frac{1}{2\pi(\sigma_1^2 \cdot \sigma_2^2 - 0 \cdot 0)^{1/2}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix} \right), \end{aligned}$$

where we have relied on the explicit formula for the determinant of a 2×2 matrix[3], and the fact that the inverse of a diagonal matrix is simply found by taking the reciprocal of each diagonal entry. Continuing,

$$\begin{aligned} p(x;\mu,\Sigma) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2}(x_1-\mu_1) \\ \frac{1}{\sigma_2^2}(x_2-\mu_2) \end{bmatrix} \right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2-\mu_2)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2 \right). \end{aligned}$$

The last equation we recognize to simply be the product of two independent Gaussian densities, one with mean μ_1 and variance σ_1^2, and the other with mean μ_2 and variance σ_2^2.

More generally, one can show that an n-dimensional Gaussian with mean $\mu \in \mathbf{R}^n$ and diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ is the same as a collection of n independent Gaussian random variables with mean μ_i and variance σ_i^2, respectively.

4 Isocontours

Another way to understand a multivariate Gaussian conceptually is to understand the shape of its **isocontours**. For a function $f : \mathbf{R}^2 \to \mathbf{R}$, an isocontour is a set of the form

$$\left\{ x \in \mathbf{R}^2 : f(x) = c \right\}.$$

for some $c \in \mathbf{R}$.[4]

[3] Namely, $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$.

[4] Isocontours are often also known as **level curves**. More generally, a **level set** of a function $f : \mathbf{R}^n \to \mathbf{R}$, is a set of the form $\left\{ x \in \mathbf{R}^2 : f(x) = c \right\}$ for some $c \in \mathbf{R}$.

4.1 Shape of isocontours

What do the isocontours of a multivariate Gaussian look like? As before, let's consider the case where $n = 2$, and Σ is diagonal, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

As we showed in the last section,

$$p(x; \mu, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right). \tag{4}$$

Now, let's consider the level set consisting of all points where $p(x; \mu, \Sigma) = c$ for some constant $c \in \mathbf{R}$. In particular, consider the set of all $x_1, x_2 \in \mathbf{R}$ such that

$$\begin{aligned} c &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\ 2\pi c\sigma_1\sigma_2 &= \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \right) \\ \log(2\pi c\sigma_1\sigma_2) &= -\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \\ \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right) &= \frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2 + \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2 \\ 1 &= \frac{(x_1 - \mu_1)^2}{2\sigma_1^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)} + \frac{(x_2 - \mu_2)^2}{2\sigma_2^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)}. \end{aligned}$$

Defining

$$r_1 = \sqrt{2\sigma_1^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)} \qquad r_2 = \sqrt{2\sigma_2^2 \log\left(\frac{1}{2\pi c\sigma_1\sigma_2} \right)},$$

it follows that

$$1 = \left(\frac{x_1 - \mu_1}{r_1} \right)^2 + \left(\frac{x_2 - \mu_2}{r_2} \right)^2. \tag{5}$$

Equation (5) should be familiar to you from high school analytic geometry: it is the equation of an **axis-aligned ellipse**, with center (μ_1, μ_2), where the x_1 axis has length $2r_1$ and the x_2 axis has length $2r_2$!

4.2 Length of axes

To get a better understanding of how the shape of the level curves vary as a function of the variances of the multivariate Gaussian distribution, suppose that we are interested in

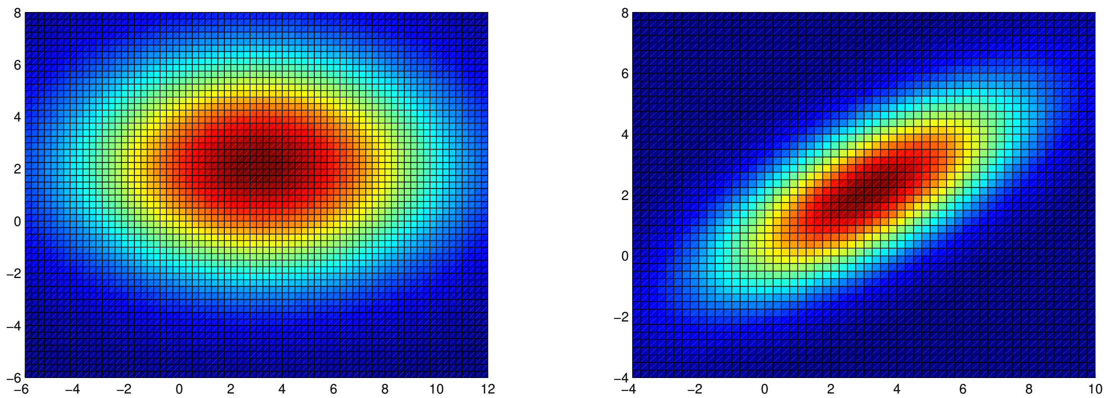

Figure 2:
The figure on the left shows a heatmap indicating values of the density function for an axis-aligned multivariate Gaussian with mean $\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and diagonal covariance matrix $\Sigma = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$. Notice that the Gaussian is centered at $(3, 2)$, and that the isocontours are all elliptically shaped with major/minor axis lengths in a 5:3 ratio. The figure on the right shows a heatmap indicating values of the density function for a non axis-aligned multivariate Gaussian with mean $\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and covariance matrix $\Sigma = \begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}$. Here, the ellipses are again centered at $(3, 2)$, but now the major and minor axes have been rotated via a linear transformation.

the values of r_1 and r_2 at which c is equal to a fraction $1/e$ of the peak height of Gaussian density.

First, observe that maximum of Equation (4) occurs where $x_1 = \mu_1$ and $x_2 = \mu_2$. Substituting these values into Equation (4), we see that the peak height of the Gaussian density is $\frac{1}{2\pi\sigma_1\sigma_2}$.

Second, we substitute $c = \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)$ into the equations for r_1 and r_2 to obtain

$$r_1 = \sqrt{2\sigma_1^2 \log\left(\frac{1}{2\pi\sigma_1\sigma_2 \cdot \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)}\right)} = \sigma_1\sqrt{2}$$

$$r_2 = \sqrt{2\sigma_2^2 \log\left(\frac{1}{2\pi\sigma_1\sigma_2 \cdot \frac{1}{e}\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)}\right)} = \sigma_2\sqrt{2}.$$

From this, it follows that the axis length needed to reach a fraction $1/e$ of the peak height of the Gaussian density in the ith dimension grows in proportion to the standard deviation σ_i. Intuitively, this again makes sense: the smaller the variance of some random variable x_i, the more "tightly" peaked the Gaussian distribution in that dimension, and hence the smaller the radius r_i.

4.3 Non-diagonal case, higher dimensions

Clearly, the above derivations rely on the assumption that Σ is a diagonal matrix. However, in the non-diagonal case, it turns out that the picture is not all that different. Instead of being an axis-aligned ellipse, the isocontours turn out to be simply **rotated ellipses**. Furthermore, in the n-dimensional case, the level sets form geometrical structures known as **ellipsoids** in $\mathbf{R}^n$.

5 Linear transformation interpretation

In the last few sections, we focused primarily on providing an intuition for how multivariate Gaussians with diagonal covariance matrices behaved. In particular, we found that an n-dimensional multivariate Gaussian with diagonal covariance matrix could be viewed simply as a collection of n independent Gaussian-distributed random variables with means and variances μ_i and σ_i^2, respectvely. In this section, we dig a little deeper and provide a quantitative interpretation of multivariate Gaussians when the covariance matrix is not diagonal.

The key result of this section is the following theorem (see proof in Appendix A.2).

Theorem 1. *Let $X \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbf{R}^n$ and $\Sigma \in \mathbf{S}^n_{++}$. Then, there exists a matrix $B \in \mathbf{R}^{n\times n}$ such that if we define $Z = B^{-1}(X - \mu)$, then $Z \sim \mathcal{N}(0, I)$.*

To understand the meaning of this theorem, note that if $Z \sim \mathcal{N}(0, I)$, then using the analysis from Section 4, Z can be thought of as a collection of n independent standard normal random variables (i.e., $Z_i \sim \mathcal{N}(0, 1)$). Furthermore, if $Z = B^{-1}(X - \mu)$ then $X = BZ + \mu$ follows from simple algebra.

Consequently, the theorem states that any random variable X with a multivariate Gaussian distribution can be interpreted as the result of applying a linear transformation ($X = BZ + \mu$) to some collection of n independent standard normal random variables (Z).

Appendix A.1

Proof. We prove the first of the two equalities in (1); the proof of the other equality is similar.

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[(X_1 - \mu_1)^2] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & \cdots & E[(X_n - \mu_n)^2] \end{bmatrix} \\
&= E \begin{bmatrix} (X_1 - \mu_1)^2 & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\ \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & \cdots & (X_n - \mu_n)^2 \end{bmatrix} && (6) \\
&= E \left[\begin{bmatrix} X_1 - \mu_1 \\ \vdots \\ X_n - \mu_n \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 & \cdots & X_n - \mu_n \end{bmatrix} \right] && (7) \\
&= E\left[(X - \mu)(X - \mu)^T \right].
\end{aligned}
$$

Here, (6) follows from the fact that the expectation of a matrix is simply the matrix found by taking the componentwise expectation of each entry. Also, (7) follows from the fact that for any vector $z \in \mathbf{R}^n$,

$$
zz^T = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \cdots z_n \end{bmatrix} = \begin{bmatrix} z_1 z_1 & z_1 z_2 & \cdots & z_1 z_n \\ z_2 z_1 & z_2 z_2 & \cdots & z_2 z_n \\ \vdots & \vdots & \ddots & \vdots \\ z_n z_1 & z_n z_2 & \cdots & z_n z_n \end{bmatrix}.
$$

□

Appendix A.2

We restate the theorem below:

Theorem 1. *Let $X \sim \mathcal{N}(\mu, \Sigma)$ for some $\mu \in \mathbf{R}^n$ and $\Sigma \in \mathbf{S}^n_{++}$. Then, there exists a matrix $B \in \mathbf{R}^{n \times n}$ such that if we define $Z = B^{-1}(X - \mu)$, then $Z \sim \mathcal{N}(0, I)$.*

The derivation of this theorem requires some advanced linear algebra and probability theory and can be skipped for the purposes of this class. Our argument will consist of two parts. First, we will show that the covariance matrix Σ can be factorized as $\Sigma = BB^T$ for some invertible matrix B. Second, we will perform a "change-of-variable" from X to a different vector valued random variable Z using the relation $Z = B^{-1}(X - \mu)$.

Step 1: Factorizing the covariance matrix. Recall the following two properties of symmetric matrices from the notes on linear algebra[5]:

1. Any real symmetric matrix $A \in \mathbf{R}^{n\times n}$ can always be represented as $A = U\Lambda U^T$, where U is a full rank orthogonal matrix containing of the eigenvectors of A as its columns, and Λ is a diagonal matrix containing A's eigenvalues.

2. If A is symmetric positive definite, all its eigenvalues are positive.

Since the covariance matrix Σ is positive definite, using the first fact, we can write $\Sigma = U\Lambda U^T$ for some appropriately defined matrices U and Λ. Using the second fact, we can define $\Lambda^{1/2} \in \mathbf{R}^{n\times n}$ to be the diagonal matrix whose entries are the square roots of the corresponding entries from Λ. Since $\Lambda = \Lambda^{1/2}(\Lambda^{1/2})^T$, we have

$$\Sigma = U\Lambda U^T = U\Lambda^{1/2}(\Lambda^{1/2})^T U^T = U\Lambda^{1/2}(U\Lambda^{1/2})^T = BB^T,$$

where $B = U\Lambda^{1/2}$.[6] In this case, then $\Sigma^{-1} = B^{-T}B^{-1}$, so we can rewrite the standard formula for the density of a multivariate Gaussian as

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|BB^T|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T B^{-T}B^{-1}(x-\mu)\right). \tag{8}$$

Step 2: Change of variables. Now, define the vector-valued random variable $Z = B^{-1}(X-\mu)$. A basic formula of probability theory, which we did not introduce in the section notes on probability theory, is the "change-of-variables" formula for relating vector-valued random variables:

> Suppose that $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T \in \mathbf{R}^n$ is a vector-valued random variable with joint density function $f_X : \mathbf{R}^n \to \mathbf{R}$. If $Z = H(X) \in \mathbf{R}^n$ where H is a bijective, differentiable function, then Z has joint density $f_Z : \mathbf{R}^n \to \mathbf{R}$, where
>
> $$f_Z(z) = f_X(x) \cdot \left| \det\left(\begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \cdots & \frac{\partial x_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial z_1} & \cdots & \frac{\partial x_n}{\partial z_n} \end{bmatrix} \right) \right|.$$

Using the change-of-variable formula, one can show (after some algebra, which we'll skip) that the vector variable Z has the following joint density:

$$p_Z(z) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}z^T z\right). \tag{9}$$

The claim follows immediately. $\square$

[5]See section on "Eigenvalues and Eigenvectors of Symmetric Matrices."

[6]To show that B is invertible, it suffices to observe that U is an invertible matrix, and right-multiplying U by a diagonal matrix (with no zero diagonal entries) will rescale its columns but will not change its rank.