EECS 189 Introduction to Machine Learning
Fall 2020

Note 6

1 Multivariate Gaussians

So far in our discussion of MLE and MAP in regression, we considered a set of Gaussian random variables $Z_1, Z_2, \ldots, Z_k$, which can represent anything from the noise in data to the parameters of a model. One critical assumption we made is that these variables are independent and identically distributed. However, what about the case when these variables are dependent and/or non-identical? For example, in time series data we have the relationship

$$Z_{i+1} = rZ_i + U_i$$

where $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $-1 \leq r \leq 1$ (so that it doesn't blow up)
Here's another example: consider the "sliding window" (like echo of audio)

$$Z_i = \mathbf{\Sigma} r_j U_{i-j}$$

where $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
In general, if we can represent the random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)$ as

$$\mathbf{Z} = \mathbf{R}\mathbf{U}$$

where $\mathbf{Z} \in \mathbb{R}^n$, $\mathbf{R} \in \mathbb{R}^{n \times n}$, $\mathbf{U} \in \mathbb{R}^n$, and $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, we refer to $\mathbf{Z}$ as a **Jointly Gaussian Random Vector**. Our goal now is to derive its probability density formula.

1.1 Definition

There are three equivalent definitions of a jointly Gaussian (JG) random vector:

1. A random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)$ is JG if there exists a base random vector $\mathbf{U} = (U_1, U_2, \ldots, U_l)$ whose components are independent standard normal random variables, a transition matrix $\mathbf{R} \in \mathbb{R}^{k \times l}$, and a mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$, such that $\mathbf{Z} = \mathbf{R}\mathbf{U} + \boldsymbol{\mu}$.
2. A random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)^\top$ is JG if $\sum_{i=1}^k a_i Z_i$ is normally distributed for every $a = (a_1, a_2, \ldots, a_k)^\top \in \mathbb{R}^k$.
3. (Non-degenerate case only) A random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_k)^\top$ is JG if
$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{|\det(\mathbf{\Sigma})|}} \frac{1}{(\sqrt{2\pi})^k} e^{-\frac{1}{2}(\mathbf{Z}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{Z}-\boldsymbol{\mu})}$$
Where $\mathbf{\Sigma} = \mathbb{E}[(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^\top] = \mathbb{E}[(\mathbf{R}\mathbf{U})(\mathbf{R}\mathbf{U})^\top] = \mathbf{R}\mathbb{E}[\mathbf{U}\mathbf{U}^\top]\mathbf{R}^\top = \mathbf{R}\mathbf{I}\mathbf{R}^\top = \mathbf{R}\mathbf{R}^\top$
$\mathbf{\Sigma}$ is also called the **covariance matrix** of $\mathbf{Z}$.

Note that all of these conditions are equivalent. In this note we will start by showing a proof that (1) $\implies$ (3). We will leave it as an exercise to prove the rest of the implications needed to show that the three conditions are in fact equivalent.

1.1.1 Proving (1) $\implies$ (3)

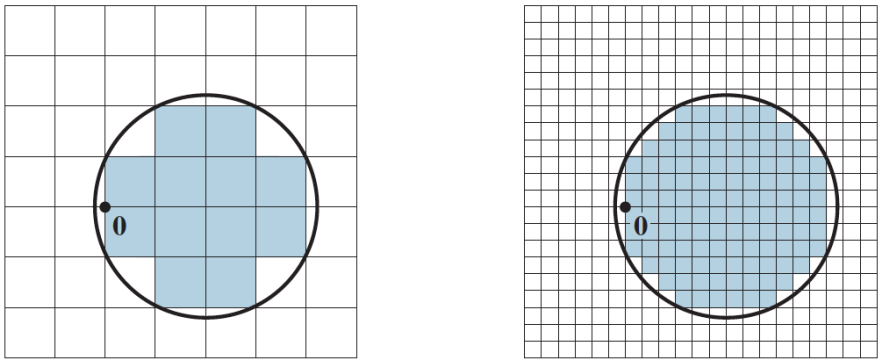
In the context of the noise problem we defined earlier, we are starting with condition (1), ie. $\mathbf{Z} = \mathbf{R}\mathbf{U}$ (in this case $k = l = n$), and we would like to derive the probability density of $\mathbf{Z}$. Note that here we removed the $\boldsymbol{\mu}$ from consideration because in machine learning we always assume that the noise has a mean of 0. We leave it as an exercise for the reader to prove the case for an arbitrary $\boldsymbol{\mu}$.

We will first start by relating the probability density function of $\mathbf{U}$ to that of $\mathbf{Z}$. Denote $f_{\mathbf{U}}(\mathbf{u})$ as the probability density for $\mathbf{U} = \mathbf{u}$, and similarly denote $f_{\mathbf{Z}}(\mathbf{z})$ as the probability density for $\mathbf{Z} = \mathbf{z}$.

One may initially believe that $f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{Z}}(\mathbf{R}\mathbf{u})$, but this is NOT true. Remember that since there is a change of variables from $\mathbf{U}$ to $\mathbf{Z}$, we must make sure to incorporate the change of variables constant, which in this case is the absolute value of the determinant of $\mathbf{R}$. Incorporating this constant, we will have the correct formula:

$$f_{\mathbf{U}}(\mathbf{u}) = |\det(\mathbf{R})| f_{\mathbf{Z}}(\mathbf{R}\mathbf{u})$$

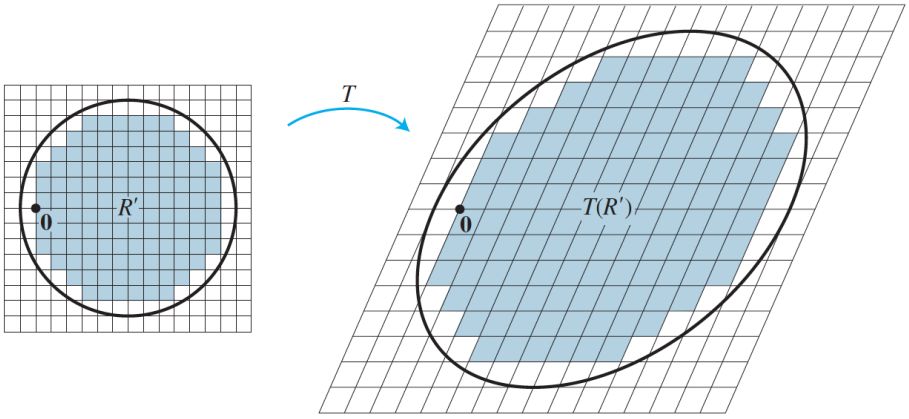
Let's see why this is true, with a simple 2D geometric explanation. Define $\mathbf{U}$ space to be the 2D space with axes U_1 and U_2. Now take any arbitrary region $\mathbf{R}'$ in $\mathbf{U}$ space (note that this $\mathbf{R}'$ is different from the matrix $\mathbf{R}$ that relates $\mathbf{U}$ to $\mathbf{Z}$). As shown in the diagram below, we have some off-centered circular region $\mathbf{R}'$ and we would like to approximate the probability that $\mathbf{U}$ takes a value in this region. We can do so by taking a Riemann sum of the density function $f_{\mathbf{U}}(.)$ over smaller and smaller squares that make up the region $\mathbf{R}'$:



Mathematically, we have that

$$P(\mathbf{U} \subseteq \mathbf{R}') = \iint_{\mathbf{R}'} f_{\mathbf{U}}(u_1, u_2)\, du_1\, du_2 \approx \sum \sum_{\mathbf{R}'} f_{\mathbf{U}}(u_1, u_2)\, \Delta u_1\, \Delta u_2$$

Now, let's apply the linear transformation $\mathbf{Z} = \mathbf{R}\mathbf{U}$, mapping the region $\mathbf{R}'$ in $\mathbf{U}$ space, to the region $T(\mathbf{R}')$ in $\mathbf{Z}$ space.



The graph on the right is now **Z** space, the 2D space with axes Z_1 and Z_2. Assuming that the matrix **R** is invertible, there is a one-to-one correspondence between points in **U** space to points in **Z** space. As we can note in the diagram above, each unit square in **U** space maps to a parallelogram in **Z** space (in higher dimensions, we would use the terms **hypercube** and **parallelepiped**). Recall the relationship between each unit hypercube and the parallelepiped it maps to:

$$\text{Area(parallelepiped)} = |\det(\mathbf{R})| \cdot \text{Area(hypercube)}$$

In this 2D example, if we denote the area of each unit square as $\Delta u_1 \Delta u_2$, and the area of each unit parallelepiped as ΔA, we say that

$$\Delta A = |\det(\mathbf{R})| \cdot \Delta u_1 \Delta u_2$$

Now let's take a Riemann sum to find the probability that **Z** takes a value in $T(\mathbf{R}')$:

$$\begin{aligned} P(\mathbf{Z} \subseteq T(\mathbf{R}')) &= \iint_{T(\mathbf{R}')} f_{\mathbf{Z}}(z_1, z_2)\, dz_1\, dz_2 \\ &\approx \sum \sum_{T(\mathbf{R}')} f_{\mathbf{Z}}(\mathbf{z})\, \Delta A \\ &= \sum \sum_{\mathbf{R}'} f_{\mathbf{Z}}(\mathbf{Ru})\, |\det(\mathbf{R})| \Delta u_1 \Delta u_2 \end{aligned}$$

Note the change of variables in the last step: we sum over the squares in **U** space, instead of parallelograms in **R** space.

So far, we have shown that (for any dimension n)

$$P(\mathbf{U} \subseteq \mathbf{R}') = \int \ldots \iint_{\mathbf{R}'} f_{\mathbf{U}}(\mathbf{u})\, du_1 du_2 \ldots du_n$$

and

$$P(\mathbf{Z} \subseteq T(\mathbf{R}')) = \int \ldots \iint_{\mathbf{R}'} f_{\mathbf{Z}}(\mathbf{Ru})\, |\det(\mathbf{R})| du_1 du_2 \ldots du_n$$

Notice that these two probabilities are equivalent! The probability that $\mathbf{U}$ takes value in $\mathbf{R}'$ must equal the probability that the transformed random vector $\mathbf{Z}$ takes a value in the transformed region $T(\mathbf{R}')$.

Therefore, we can say that

$$\begin{aligned} P(\mathbf{U} \subseteq \mathbf{R}') &= \int \dots \iint_{\mathbf{R}'} f_{\mathbf{U}}(\mathbf{u})\, du_1 du_2 \dots du_n \\ &= \int \dots \iint_{\mathbf{R}'} f_{\mathbf{Z}}(\mathbf{R}\mathbf{u})\, |\det(\mathbf{R})| du_1 du_2 \dots du_n \\ &= P(\mathbf{Z} \subseteq T(\mathbf{R}')) \end{aligned}$$

We conclude that

$$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{Z}}(\mathbf{R}\mathbf{u})\, |\det(\mathbf{R})|$$

An almost identical argument will allow us to state that

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{U}}(\mathbf{R}^{-1}\mathbf{z})\, |\det\left(\mathbf{R}^{-1}\right)| = \frac{1}{|\det(\mathbf{R})|} f_{\mathbf{U}}(\mathbf{R}^{-1}\mathbf{z})$$

Since the densities for all the U_i's are i.i.d, and $\mathbf{U} = \mathbf{R}^{-1}\mathbf{Z}$, we can write the joint density function of Z as

$$\begin{aligned} f_{\mathbf{Z}}(\mathbf{z}) &= \frac{1}{|\det(\mathbf{R})|} f_{\mathbf{U}}(\mathbf{R}^{-1}\mathbf{z}) \\ &= \frac{1}{|\det(\mathbf{R})|} \prod_{i=1}^{n} f_{U_i}((\mathbf{R}^{-1}\mathbf{z})_i) \\ &= \frac{1}{|\det(\mathbf{R})|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(\mathbf{R}^{-1}\mathbf{z})^{\top}(\mathbf{R}^{-1}\mathbf{z})} \\ &= \frac{1}{|\det(\mathbf{R})|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{z}^{\top}\mathbf{R}^{-T}\mathbf{R}^{-1}\mathbf{z}} \\ &= \frac{1}{|\det(\mathbf{R})|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{z}^{\top}(\mathbf{R}\mathbf{R}^{\top})^{-1}\mathbf{z}} \end{aligned}$$

Note that $(\mathbf{R}\mathbf{R}^{\top})^{-1}$ is simply the covariance matrix for $\mathbf{Z}$:

$$\text{Cov}[\mathbf{Z}] = \mathbb{E}[\mathbf{Z}\mathbf{Z}^{\top}] = \mathbb{E}[\mathbf{R}\mathbf{U}\mathbf{U}^{\top}\mathbf{R}^{\top}] = \mathbf{R}\mathbb{E}[\mathbf{U}\mathbf{U}^{\top}]\mathbf{R}^{\top} = \mathbf{R}\mathbf{I}\mathbf{R}^{\top} = \mathbf{R}\mathbf{R}^{\top}$$

Thus the density function of $\mathbf{Z}$ can be written as

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{|\det(\mathbf{R})|} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{z}^{\top}\mathbf{\Sigma}_Z^{-1}\mathbf{z}}$$

Furthermore, we know that

$$|\det(\mathbf{\Sigma}_Z)| = |\det\left(\mathbf{R}\mathbf{R}^{\top}\right)|$$

$$= |\det(\mathbf{R}) \cdot \det\left(\mathbf{R}^\top\right)|$$
$$= |\det(\mathbf{R}) \cdot \det(\mathbf{R})| = |\det(\mathbf{R})|^2$$

and therefore

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{\sqrt{\det(\mathbf{\Sigma}_Z)}} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{\Sigma}_Z^{-1} \mathbf{z}}$$

1.2 Estimating Gaussians from Data

For a particular multivariate Gaussian distribution $f(.)$, if we do not have the true means and covariances $\boldsymbol{\mu}, \mathbf{\Sigma}$, then our best bet is to use MLE to estimate them empirically with i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{t_i=k} \mathbf{x}_i$$
$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{t_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

Note that the above formulas are not necessarily trivial and must be formally proven using MLE. Just to present a glimpse of the process, let's prove that these formulas hold for the case where we are dealing with 1-d data points. For notation purposes, assume that $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ is the set of all training data points that belong to class k. Note that the data points are i.i.d. Our goal is to solve the following MLE problem:

$$\begin{aligned}
\hat{\mu}, \hat{\sigma}^2 &= \arg\max_{\mu,\sigma^2} \; P(x_1, x_2, ..., x_n \mid \mu, \sigma^2) \\
&= \arg\max_{\mu,\sigma^2} \; \ln\Big(P(x_1, x_2, ..., x_n \mid \mu, \sigma^2)\Big) \\
&= \arg\max_{\mu,\sigma^2} \; \sum_{i=1}^{n} \ln\Big(P(x_i \mid \mu, \sigma^2)\Big) \\
&= \arg\max_{\mu,\sigma^2} \; \sum_{i=1}^{n} -\frac{(x_i - \mu)^2}{2\sigma^2} - \ln(\sigma) - \frac{1}{2}\ln(2\pi) \\
&= \arg\min_{\mu,\sigma^2} \; \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma)
\end{aligned}$$

Note that the objective above is not jointly convex, so we cannot simply take derivatives and set them to 0! Instead, we decompose the minimization over σ^2 and μ into a nested optimization problem:

$$\min_{\mu,\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma) = \min_{\sigma^2} \min_{\mu} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} + \ln(\sigma)$$

The optimization problem has been decomposed into an inner problem that optimizes for μ given a fixed σ^2, and an outer problem that optimizes for σ^2 given the optimal value $\hat{\mu}$. Let's first solve

the inner optimization problem. Given a fixed σ^2, the objective is convex in μ, so we can simply take a partial derivative w.r.t μ and set it equal to 0:

$$\frac{\partial}{\partial \mu}\Big(\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} + \ln(\sigma)\Big) = \sum_{i=1}^{n} \frac{-(x_i-\mu)}{\sigma^2} = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Having solved the inner optimization problem, we now have that

$$\min_{\sigma^2}\min_{\mu} \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} + \ln(\sigma) = \min_{\sigma^2} \sum_{i=1}^{n} \frac{(x_i-\hat{\mu})^2}{2\sigma^2} + \ln(\sigma)$$

Note that this objective is not convex in σ, so we must instead find the critical point of the objective that minimizes the objective. Assuming that $\sigma \geq 0$, the critical points are:

- $\sigma = 0$: assuming that not all of the points x_i are equal to $\hat{\mu}$, there are two terms that are at odds with each other: a $1/\sigma^2$ term that blows off to ∞, and a $\ln(\sigma)$ term that blows off to $-\infty$ as $\sigma \to 0$. Note that the $1/\sigma^2$ term blows off at a faster rate, so we conclude that

$$\lim_{\sigma \to 0} \sum_{i=1}^{n} \frac{(x_i-\hat{\mu})^2}{2\sigma^2} + \ln(\sigma) = \infty$$

- $\sigma = \infty$: this case does not lead to the solution, because it gives a maximum, not a minimum.

$$\lim_{\sigma \to \infty} \sum_{i=1}^{n} \frac{(x_i-\hat{\mu})^2}{2\sigma^2} + \ln(\sigma) = \infty$$

- Points at which the derivative w.r.t σ is 0

$$\frac{\partial}{\partial \sigma}\Big(\sum_{i=1}^{n} \frac{(x_i-\hat{\mu})^2}{2\sigma^2} + \ln(\sigma)\Big) = \sum_{i=1}^{n} -\frac{(x_i-\hat{\mu})^2}{\sigma^3} + \frac{1}{\sigma} = 0 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

We conclude that the optimal point is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

1.3 Isocontours

Let's try to understand in detail how to visualize a multivariate Gaussian distribution. For simplicity, let's consider a zero-mean Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, which just leaves us with the covariance matrix $\mathbf{\Sigma}$. Since $\mathbf{\Sigma}$ is a symmetric, positive semidefinite matrix, we can decompose it by the spectral theorem into $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where the columns of $\mathbf{V}$ form an orthonormal basis in $\mathbb{R}^d$, and $\mathbf{\Lambda}$ is a diagonal matrix with real, non-negative values. We wish to find its **level set**

$$f(\mathbf{x}) = k$$

or simply the set of all points $\mathbf{x}$ such that the probability density $f(\mathbf{x})$ evaluates to a fixed constant k. This is equivalent to the level set $\ln(f(\mathbf{x})) = \ln(k)$ which further reduces to

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = c$$

for some constant c. Without loss of generality, assume that this constant is 1. The level set $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = 1$ is an ellipsoid with axes $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$, with lengths $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_d}$, respectively. Each axis of the ellipsoid is the vector $\sqrt{\lambda_i}\mathbf{v}_i$, and we can verify that

$$(\sqrt{\lambda_i}\mathbf{v}_i)^T\boldsymbol{\Sigma}^{-1}(\sqrt{\lambda_i}\mathbf{v}_i) = \lambda_i\mathbf{v}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{v}_i = \lambda_i\mathbf{v}_i^T(\boldsymbol{\Sigma}^{-1}\mathbf{v}_i) = \lambda_i\mathbf{v}_i^T(\lambda_i^{-1}\mathbf{v}_i) = \mathbf{v}_i^T\mathbf{v}_i = 1$$

The entries of $\boldsymbol{\Lambda}$ dictate how elongated or shrunk the distribution is along each direction. In the case of **isotropic** distributions, the entries of $\boldsymbol{\Lambda}$ are all identical, meaning the the axes of the ellipsoid form a circle. In the case of **anisotropic** distributions, the entries of $\boldsymbol{\Lambda}$ are not necessarily identical, meaning that the resulting ellipsoid may be elongated/shrunken and also rotated.

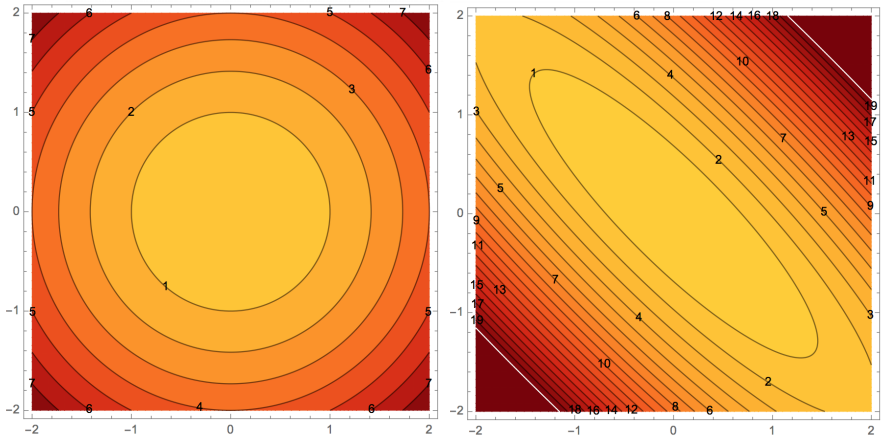


Figure 1: Isotropic (left) vs Anisotropic (right) contours are ellipsoids with axes $\sqrt{\lambda_i}\mathbf{v}_i$. Images courtesy Professor Shewchuk's notes

1.4 Properties

Let's state some well-known properties of Multivariate Gaussians. Given a JG random vector $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{Z}, \boldsymbol{\Sigma}_\mathbf{Z})$, the linear transformation $\mathbf{AZ}$ (where $\mathbf{A}$ is an appropriately dimensioned constant matrix) is also JG:

$$\mathbf{AZ} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_\mathbf{Z}, \mathbf{A}\boldsymbol{\Sigma}_\mathbf{Z}\mathbf{A}^\top)$$

We can derive the mean and covariance of $\mathbf{AZ}$ using the linearity of expectations:

$$\boldsymbol{\mu}_\mathbf{AZ} = \mathbb{E}[\mathbf{AZ}] = \mathbf{A}\mathbb{E}[\mathbf{Z}] = \mathbf{A}\boldsymbol{\mu}_\mathbf{Z}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_\mathbf{AZ} &= \mathbb{E}[(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])(\mathbf{AZ} - \mathbb{E}[\mathbf{AZ}])^\top] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top\mathbf{A}^\top] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])^\top]\mathbf{A}^\top\end{aligned}$$

$$= \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{Z}}\mathbf{A}^{\top}$$

Note that the statements above did not rely on the fact that $\mathbf{Z}$ is JG, so this reasoning applies to all random vectors. We know that $\mathbf{AZ}$ is JG itself, because it can be expressed as a linear transformation of i.i.d. Gaussians: $\mathbf{AZ} = \mathbf{ARU}$.

Now suppose that we have the partition $\mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}$ whose distribution is given by $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}})$ and

$$\boldsymbol{\mu}_{\mathbf{Z}} = \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{X}}\\ \boldsymbol{\mu}_{\mathbf{Y}}\end{bmatrix}, \boldsymbol{\Sigma}_{\mathbf{Z}} = \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}}\\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}}\end{bmatrix}$$

It turns out that the **marginal distribution** of the individual random vector $\mathbf{X}$ (and $\mathbf{Y}$) is JG:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{XX}})$$

However, the converse is not necessarily true: if $\mathbf{X}$ and $\mathbf{Y}$ are each individually JG, it is not necessarily the case that $\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}$ is JG! To see why, let's suppose that $\mathbf{X}$ and $\mathbf{Y}$ are individually JG. Thus, we can express each as a linear transformation of i.i.d. Gaussian random variables:

$$\mathbf{X} = \mathbf{R}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}, \mathbf{Y} = \mathbf{R}_{\mathbf{Y}}\mathbf{U}_{\mathbf{Y}}$$

we would expect that the expression for the joint distribution would be JG:

$$\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} = \begin{bmatrix}\mathbf{R}_{\mathbf{X}} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{\mathbf{Y}}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{\mathbf{X}}\\ \mathbf{U}_{\mathbf{Y}}\end{bmatrix}$$

However, since we cannot guarantee that the entries of $\mathbf{U}_{\mathbf{X}}$ are independently distributed from the entries of $\mathbf{U}_{\mathbf{Y}}$, we cannot conclude that the joint distribution is JG. If the entries are independently distributed, then we would be able to conclude that the joint distribution is JG.

Let's now transition back to our discussion of $\mathbf{Z}$. The **conditional distribution** of $\mathbf{X}$ given $\mathbf{Y}$ (and vice versa) is also JG:

$$\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}), \boldsymbol{\Sigma}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}})$$

If $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated (that is, if $\boldsymbol{\Sigma}_{\mathbf{XY}} = \boldsymbol{\Sigma}_{\mathbf{YX}} = \mathbf{0}$), we can say that they are independent. Namely, the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ does not depend on $\mathbf{Y}$:

$$\mathbf{X}|\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \mathbf{0}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}}), \boldsymbol{\Sigma}_{\mathbf{XX}} - \mathbf{0}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{0}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{XX}})$$

This also follows from the multivariate Gaussian pdf:

$$\begin{aligned} f_{\mathbf{Z}}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} &= \frac{1}{(\sqrt{2\pi})^{n}}\left|\begin{matrix}\boldsymbol{\Sigma}_{\mathbf{XX}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{YY}}\end{matrix}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x} & \mathbf{y}\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{XX}} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{YY}}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right)\\ &= \frac{1}{(\sqrt{2\pi})^{n_x}}\left|\boldsymbol{\Sigma}_{\mathbf{XX}}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{x}^{\top}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\mathbf{x}\right)\cdot\frac{1}{(\sqrt{2\pi})^{n_y}}\left|\boldsymbol{\Sigma}_{\mathbf{YY}}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{y}^{\top}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\mathbf{y}\right)\end{aligned}$$

$$= f_{\mathbf{X}}(\mathbf{x}) \cdot f_{\mathbf{Y}}(\mathbf{y})$$

Note the significance of this statement. Given any two general random vectors, we cannot necessarily say "if they are uncorrelated, then they are independent". However in the case of random vectors from the same JG joint distribution, we can make this claim.