# MVN Example

Eric Ingram

2024-10-23

erimingram.github.io

## MVN Examples

Here are two examples of calculating marginal and conditional distributions from a Multivariate Normal. If you are reading this as a AI/CS/DSC 391L student, you will not have to calculate inverses by hand.

### Question 1:

Assume that the joint distribution of the average daily temperatures on three consecutive days, $X_1, X_2, X_3$ in Austin, TX in September is Multivariate Normal with mean and covariance

$$\mu = \begin{pmatrix} 94 \\ 93 \\ 95 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} 7 & 1 & .3 \\ 1 & 7 & 2 \\ .3 & 2 & 7 \end{pmatrix}$$

The average temperature today was 92 degrees and yesterday was 89 degrees.

Find the probability that the average temperature tomorrow will be greater than 98 degrees.

How does this probability compare to the probability of the average temperature tomorrow being greater than 98 degrees if you didn't know the average temperature today and yesterday?

**Answer:**

**Conditional:** The first question is asking us to find a probability based on a conditional distribution. We have $\mathbf{X} = \begin{pmatrix} 89 \\ 92 \\ X_3 \end{pmatrix}$, so we need to find the distribution of $X_3 | X_1 = 89, X_2 = 92$.

Remember for conditional distribution stuff, we need to partition everything into two blocks: one with the known variables and one with the unknown variables. I will call the known variables (what we are conditioning on) $\mathbf{Y}$ and the unknown variables $\mathbf{Z}$.

Here, the way we have things set up, we have $\mathbf{\Sigma}_{\text{reordered}} = \begin{pmatrix} \mathbf{\Sigma}_{YY} & \mathbf{\Sigma}_{YZ} \\ \mathbf{\Sigma}_{ZY} & \mathbf{\Sigma}_{ZZ} \end{pmatrix}$. It is usually convention to have the known variables in the top left.

So, partitioning our mean matrix we get that $\mu_Y = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 94 \\ 93 \end{pmatrix}$ and $\mu_Z = \mu_3 = 95$.

When we partition our covariance matrix, we end up with the covariance among known variables $\mathbf{\Sigma}_{YY} = \begin{pmatrix} 7 & 1 \\ 1 & 7 \end{pmatrix}$, the covariance between known and unknown variables $\mathbf{\Sigma}_{YZ} = \begin{pmatrix} .3 \\ 2 \end{pmatrix}$, $\mathbf{\Sigma}_{ZY} = \mathbf{\Sigma}_{YZ}^T = \begin{pmatrix} .3 & 2 \end{pmatrix}$, and the variance of the unknown variable $\sigma_{ZZ} = \sigma_{33} = 7$. Note that I used $\sigma$ here instead of $\mathbf{\Sigma}$: just notation to denote that it is a scalar. It is fine if you represent it as a 1 x 1 matrix.

We also have our known data vector $y = \begin{pmatrix} 89 \\ 92 \end{pmatrix}$

We know that the conditional mean $\mu_{Z|Y} = \mu_Z + \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}(y - \mu_Y)$.

So, we need to find the inverse of $\mathbf{\Sigma}_{YY}$. I'm going to do this by hand for completeness, but most people would just do this with a calculator or code. If you are confused on how I did this, review your basic linear algebra. $\mathbf{\Sigma}_{YY}^{-1} = \frac{1}{\det(\mathbf{\Sigma}_{YY})}\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix} = \frac{1}{48}\begin{pmatrix} 7 & -1 \\ -1 & 7 \end{pmatrix}$.

Now, let's calculate the conditional mean piece by piece.

$(y - \mu_Y) = \begin{pmatrix} 89 \\ 92 \end{pmatrix} - \begin{pmatrix} 94 \\ 93 \end{pmatrix} = \begin{pmatrix} -5 \\ -1 \end{pmatrix}$.

$\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1} = \begin{pmatrix} .3 & 2 \end{pmatrix} * \frac{1}{48}\begin{pmatrix} 7 & -1 \\ -1 & 7 \end{pmatrix} = \frac{1}{48}\begin{pmatrix} .1 & 13.7 \end{pmatrix}$.

Then, $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}(y - \mu_Y) = \frac{1}{48}\begin{pmatrix} .1 & 13.7 \end{pmatrix} * \begin{pmatrix} -5 \\ -1 \end{pmatrix} = -.2958$.

So, $\mu_{Z|Y} = 95 - .2958 = 94.704$.

Remember that the conditional variance $\sigma_{ZZ|Y} = \sigma_{ZZ} - \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ}$.

We have all the matrices already from earlier! Let's compute $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ} = \begin{pmatrix} .3 & 2 \end{pmatrix} * \frac{1}{48}\begin{pmatrix} 7 & -1 \\ -1 & 7 \end{pmatrix} * \begin{pmatrix} .3 \\ 2 \end{pmatrix} = \frac{1}{48}\begin{pmatrix} .1 & 13.7 \end{pmatrix} * \begin{pmatrix} .3 \\ 2 \end{pmatrix} = .571$.

Then, $\sigma_{ZZ|Y} = 7 - .571 = 6.429$

Now we have the conditional distribution! $X_3|X_1 = 89, X_2 = 92 \sim N(\mu_{Z|Y}, \sigma_{ZZ|Y}) = N(94.704, 6.429)$.

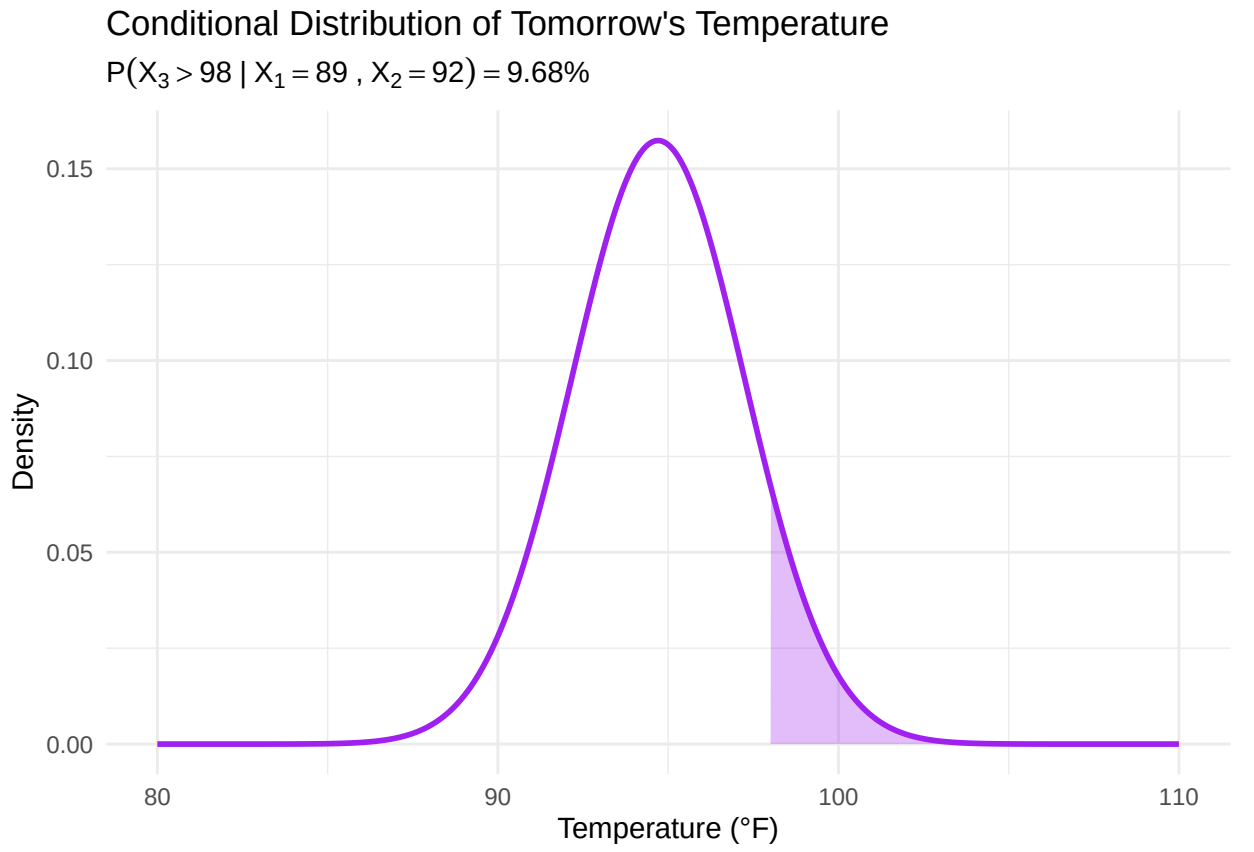
To find $P(X_3 > 98|X_1 = 89, X_2 = 92)$, we need to use the normal CDF. I won't bother doing this by hand, and use code to find this. But, mathematically, this is $\int_{98}^{\infty} f_{X_3|X_1=89,X_2=92}(x)dx$, where $f_{X_3|X_1=89,X_2=92}(x)$ is the PDF of the conditional distribution we found above.

In practice, we would probably use a Z-score to approximate this: $Z = \frac{X_3 - \mu_{X_3|X_1=89,X_2=92}}{\sigma_{X_3|X_1=89,X_2=92}}$. So, $P(Z > 98|X_1 = 89, X_2 = 92) = P(Z > \frac{98-94.704}{\sqrt{6.429}})$.

As you can see below, we get that $P(X_3 > 98|X_1 = 89, X_2 = 92) = .0968$. So, the probability of the temperature being higher than 98 degrees tomorrow given that it was 89 degrees yesterday and 92 degrees today is 9.68%. Not very high at all!

```
prob_conditional = 1 - pnorm(98, mean = 94.704, sd = sqrt(6.429))

x_values <- seq(80, 110, length.out = 1000)

density_conditional <- dnorm(x_values, mean = 94.704, sd = sqrt(6.429))

data_conditional <- data.frame(Temperature = x_values, Density = density_conditional)

ggplot(data_conditional, aes(x = Temperature, y = Density)) +
  geom_line(color = "purple", linewidth = 1) +
  geom_area(data = subset(data_conditional, Temperature > 98),
            aes(y = Density), fill = "purple", alpha = 0.3) +
  labs(
    title = "Conditional Distribution of Tomorrow's Temperature",
    subtitle = bquote(
      P(X[3] > 98 ~ "|" ~ X[1] == 89 ~
```

```
        "," ~ X[2] == 92) == .(round(prob_conditional * 100, 2)) * "%")
  ) +
  xlab("Temperature (°F)") +
  ylab("Density") +
  theme_minimal()
```




**Marginal** The next part of the question asks us to find the average temperature tomorrow being greater than 98 degrees if you didn't know the average temperature today and yesterday.

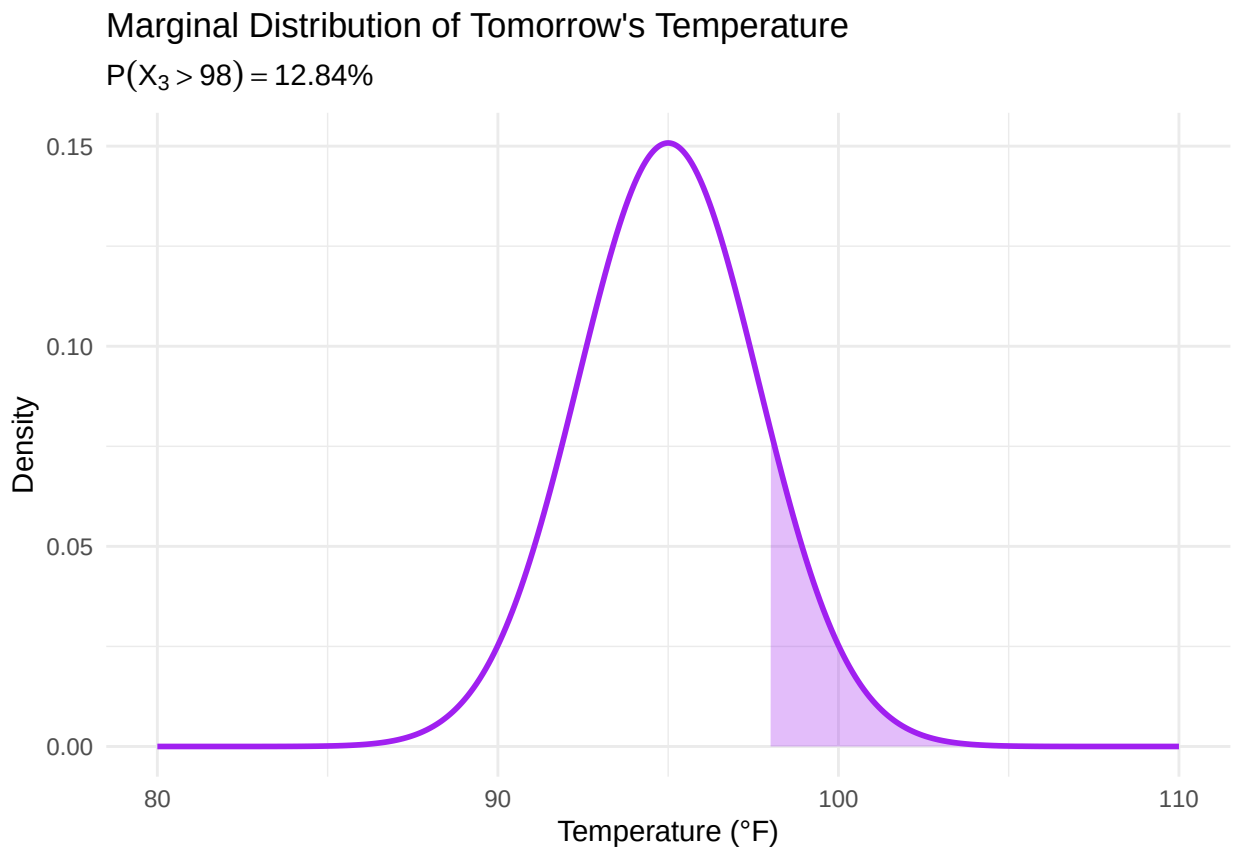
This is asking us to find the marginal probability for the temperature tomorrow: $P(X_3 > 98)$.

Remember that in a mulitvariate normal, the marginal distribution for one of the variables is just the individual normal for that variable. So, in this case, $X_3 \sim N(\mu_3, \sigma_{33}^2) = N(95, 7)$.

Then, we would just calculate the probability by using the CDF for this Normal, similar to how we calculated it above.

```
prob_marginal = 1 - pnorm(98, mean = 95, sd = sqrt(7))

x_values <- seq(80, 110, length.out = 1000)

density_marginal <- dnorm(x_values, mean = 95, sd = sqrt(7))

data_marginal <- data.frame(Temperature = x_values, Density = density_marginal)

ggplot(data_marginal, aes(x = Temperature, y = Density)) +
  geom_line(color = "purple", linewidth = 1) +
```

```r
geom_area(data = subset(data_marginal, Temperature > 98),
          aes(y = Density), fill = "purple", alpha = 0.3) +
labs(
  title = "Marginal Distribution of Tomorrow's Temperature",
  subtitle = bquote(
    P(X[3] > 98) == .(round(prob_marginal * 100, 2)) * "%")
) +
xlab("Temperature (°F)") +
ylab("Density") +
theme_minimal()
```




So overall, we can see that we have a higher probability of seeing a temperature higher than 98 degrees tomorrow if we don't account for yesterday's and today's temperature (12.84%). This should make intuitive sense: yesterday's and today's temperature were lower than the mean for $X_1$ and $X_2$ which leads to the probability being lower in the conditional case (9.68%).

## Question 2:

Suppose we have four RVs $X_1, X_2, X_3, X_4$ that follow a multivariate normal distribution: $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim MVN(\mu, \mathbf{\Sigma})$.

We have that $\mu = \begin{pmatrix} 10 \\ 20 \\ 30 \\ 40 \end{pmatrix}$ and $\mathbf{\Sigma} = \begin{pmatrix} 16 & 4 & 2 & 1 \\ 4 & 25 & 5 & 2 \\ 2 & 5 & 36 & 6 \\ 1 & 2 & 6 & 49 \end{pmatrix}$

Given $X_2 = 22, X_4 = 38$, calculate the conditional distribution of $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ given $X_2$ and $X_4$. Calculate the conditional means, variances, and covariances.

**Answer:**

We should start by doing what we did in the previous question: splitting this up into two blocks.

We need one block to have the variables we know and the lower one block to have the variables we don't know. In this case, this will require reordering the variables since the variables that need to go next to each other aren't next to each other. Remember that it is convention to have the known variables in the top left.

So, after reordering, we will have $\begin{pmatrix} X_2 \\ X_4 \\ X_1 \\ X_3 \end{pmatrix} \sim N(\mu_{\text{reordered}}, \mathbf{\Sigma}_{\text{reordered}})$, where $\mu_{reordered} = \begin{pmatrix} 20 \\ 40 \\ 10 \\ 30 \end{pmatrix}$ and $\mathbf{\Sigma}_{\text{reordered}} = \begin{pmatrix} 25 & 2 & 4 & 5 \\ 2 & 49 & 1 & 6 \\ 4 & 1 & 16 & 2 \\ 5 & 6 & 2 & 36 \end{pmatrix}$. Pause here and make sure that this reordering makes complete sense to you.

Then, let's partition this just like we did before.

Let's let $\mathbf{Z} = \begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ with $\mu_Z = \begin{pmatrix} 10 \\ 30 \end{pmatrix}$. This is our block that represents our variables of interest.

Similarly, let's let $\mathbf{Y} = \begin{pmatrix} X_2 \\ X_4 \end{pmatrix}$ with $\mu_Y = \begin{pmatrix} 20 \\ 40 \end{pmatrix}$ This is our block that represents our known variables that we are conditioning on.

Now, let's define our partitioned covariance matrices:

$\mathbf{\Sigma}_{\text{reordered}} = \begin{pmatrix} \mathbf{\Sigma}_{YY} & \mathbf{\Sigma}_{YZ} \\ \mathbf{\Sigma}_{ZY} & \mathbf{\Sigma}_{ZZ} \end{pmatrix}$,

Where $\mathbf{\Sigma}_{YY} = \begin{pmatrix} 25 & 2 \\ 2 & 49 \end{pmatrix}$

Where $\mathbf{\Sigma}_{YZ} = \begin{pmatrix} 4 & 5 \\ 1 & 6 \end{pmatrix}$

Where $\mathbf{\Sigma}_{ZY} = \mathbf{\Sigma}_{YZ}^T = \begin{pmatrix} 4 & 1 \\ 5 & 6 \end{pmatrix}$,

Where $\mathbf{\Sigma}_{ZZ} = \begin{pmatrix} 16 & 2 \\ 2 & 36 \end{pmatrix}$,

Once again, double check your understanding. Do you understand where these matrices are coming from?

**Conditional Mean** Now, let's do the conditional mean.

Remember that $\mu_{Z|Y} = \mu_Z + \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}(y - \mu_Y)$.

First, let's calculate $(y - \mu_Y)$. Given our observed data $y = \begin{pmatrix} x_2 \\ x_4 \end{pmatrix} = \begin{pmatrix} 22 \\ 38 \end{pmatrix}$, $(y - \mu_Y) = \begin{pmatrix} 22 - 20 \\ 38 - 40 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}$

Next, we need to calculate $\mathbf{\Sigma}_{YY}^{-1}$. I will leave this up to a calculator because it is extremely tedious by hand: $\mathbf{\Sigma}_{YY}^{-1} = \frac{1}{1221}\begin{pmatrix} 49 & -2 \\ -2 & 25 \end{pmatrix}$. Then, let's calculate $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1} = \begin{pmatrix} 4 & 1 \\ 5 & 6 \end{pmatrix} * \frac{1}{1221}\begin{pmatrix} 49 & -2 \\ -2 & 25 \end{pmatrix} = \frac{1}{1221}\begin{pmatrix} 194 & 17 \\ 233 & 140 \end{pmatrix}$.

Then, let's calculate $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}(y - \mu_Y) = \frac{1}{1221}\begin{pmatrix} 194 & 17 \\ 233 & 140 \end{pmatrix} * \begin{pmatrix} 2 \\ -2 \end{pmatrix} = \frac{1}{1221}\begin{pmatrix} 354 \\ 186 \end{pmatrix} = \begin{pmatrix} .2899 \\ .1523 \end{pmatrix}$.

Finally, $\mu_{Z|Y} = \mu_Z + \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}(y - \mu_Y) = \begin{pmatrix} 10 \\ 30 \end{pmatrix} + \begin{pmatrix} .2899 \\ .1523 \end{pmatrix} = \begin{pmatrix} 10.2899 \\ 30.1523 \end{pmatrix}$.

Phew! That was a lot. Now, onto the conditional covariance matrix.

**Conditional Covariance Matrix** Remember that the conditional covariance matrix is given by $\mathbf{\Sigma}_{Z|Y} = \mathbf{\Sigma}_{ZZ} - \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ}$.

Let's first compute $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ}$. We already have some of this done from earlier: $\mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ} = \frac{1}{1221}\begin{pmatrix} 194 & 17 \\ 233 & 140 \end{pmatrix} * \begin{pmatrix} 4 & 5 \\ 1 & 6 \end{pmatrix} = \frac{1}{1221}\begin{pmatrix} 793 & 1072 \\ 1072 & 2005 \end{pmatrix}$.

So, $\mathbf{\Sigma}_{Z|Y} = \mathbf{\Sigma}_{ZZ} - \mathbf{\Sigma}_{ZY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YZ} = \begin{pmatrix} 16 & 2 \\ 2 & 36 \end{pmatrix} - \frac{1}{1221}\begin{pmatrix} 793 & 1072 \\ 1072 & 2005 \end{pmatrix} = \begin{pmatrix} 15.3505 & 1.12203 \\ 1.12203 & 34.3579 \end{pmatrix}$.

This is our conditional covariance matrix.

**Alternative Approach for Conditional Covariance Matrix** Given $\mathbf{\Sigma}_{\text{reordered}} = \begin{pmatrix} 25 & 2 & 4 & 5 \\ 2 & 49 & 1 & 6 \\ 4 & 1 & 16 & 2 \\ 5 & 6 & 2 & 36 \end{pmatrix}$, let's calculate the precision matrix, $\mathbf{Q} = \mathbf{\Sigma}_{\text{reordered}}^{-1}$.

$$\mathbf{Q} = \begin{pmatrix} 0.0427126 & -0.000899705 & -0.00996835 & -0.00522856 \\ -0.000899705 & 0.0208644 & -0.000664661 & -0.00331552 \\ -0.00996835 & -0.000664661 & 0.0653002 & -0.00213252 \\ -0.00522856 & -0.00331552 & -0.00213252 & 0.029175 \end{pmatrix}$$

We can partition $\mathbf{Q}$ in the same way that we partitioned $\mathbf{\Sigma}_{\text{reordered}}$.

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{YY} & \mathbf{Q}_{YZ} \\ \mathbf{Q}_{ZY} & \mathbf{Q}_{ZZ} \end{pmatrix}$$

It is a fact that $\mathbf{\Sigma}_{Z|Y} = \mathbf{Q}_{ZZ}^{-1}$. In other words, the conditional covariance matrix $\mathbf{\Sigma}_{Z|Y}$ is obtained by inverting the block of the precision matrix corresponding to the variables $Z$

So, $\mathbf{Q}_{ZZ} = \begin{pmatrix} 0.0653002 & -0.00213252 \\ -0.00213252 & 0.029175 \end{pmatrix}$.

Taking the inverse of this yields $\mathbf{Q}_{ZZ}^{-1} = \begin{pmatrix} 15.3505 & 1.12203 \\ 1.12203 & 34.3579 \end{pmatrix} = \mathbf{\Sigma}_{Z|Y}$

Cool, right?!!

What else could we have possibly done with $\mathbf{Q}$?

The elements of $\mathbf{Q}$ are known as partial covariances. A partial covariance measures the strength of the linear relationship between two variables **given all other variables**.

So, for example, we have above that $q_{12} = -0.000899705$. Remember that after reordering this cell corresponds to $X_2$ and $X_4$ in the reordered matrix. How we would interpret this is that after controlling for all other variables (so $X_1$ and $X_3$), $X_2$ and $X_4$ have a very weak linear relationship. So once we account for the influence of $X_1$ and $X_3$, knowing $X_2$ doesn't tell you much about $X_4$ and vice versa.

For this class, the only other thing you need to care about is that if $q_{ij} = 0$, variables $X_i$ and $X_j$ are conditionally independent **given all other variables**.

**Final Results** So, our conditional distribution is $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} | X_2 = 22, X_4 = 38 \sim MVN(\begin{pmatrix} 10.2898 \\ 30.1523 \end{pmatrix}, \begin{pmatrix} 15.3505 & 1.12203 \\ 1.12203 & 34.3579 \end{pmatrix})$

Our conditional covariances are the off-diagonal elements of the covariance matrix. So, $Cov(X_1, X_3 | X_2 = 22, X_4 = 38) = 1.12203$.

For completeness, if we take the covariance of a variable that we are conditioning on with one that we aren't conditioning on, such as $Cov(X_1, X_2 | X_2 = 22, X_4 = 38)$, it is 0. Why? Just do the basic covariance algebra: $Cov(X_1, X_2 | X_2, X_4) = Cov(X_1, 22 | X_2 = 22, X_4 = 38) = 0$ since the covariance of a variable with a constant is 0.

What about conditional independence? For conditional independence, the conditional covariance between the variables has to be equal to 0. Here, $Cov(X_1, X_3 | X_2 = 22, X_4 = 38) = 1.12203 \neq 0$, so $X_1$ and $X_3$ are not conditionally independent.