Homework 1: Neo Lee

Introduction to Time Series, Fall 2023

Due Tuesday September 5 at 5pm

The total number of points possible for this homework is 41. The number of points for each question is written below, and questions marked as "bonus" are optional. Submit the **knitted html file** from this Rmd to Gradescope.

If you collaborated with anybody for this homework, put their names here:

Correlation and independence

1. (3 pts) Give an example to show that two random variables can be uncorrelated but not independent. You must explicitly prove that they are uncorrelated but not independent (for the latter, you may invoke any property that you know is equivalent to independence).

SOLUTION GOES HERE

Consider a uniform random variable X over the space of $\{-1, 0, 1\}$, and another random variable $Y = |X|$. Then, $Cov(X, Y) = E[XY] - E[X]E[Y] = (1 \times P(X = 1, Y = 1) - 1 \times P(X = -1, Y = 1) + 0 \times P(X = 0, Y = 0)) - 0 \times \frac{2}{3} = 0$. Hence, their correlation is also 0. However, we know X, Y are definitely not independent, because we have explicitly defined $Y = |X|$. We can also check by looking at $P(Y = 1|X = -1) = 1 \neq P(Y = 1) = \frac{2}{3}$.

2. (2 pts) If (X, Y) has a multivariate Gaussian distribution, and X, Y are uncorrelated: $\text{Cov}(X, Y) = 0$, then show that X, Y are independent.

SOLUTION GOES HERE

We just have to show that $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$. Now we evaluate the joint distribution

$$\begin{aligned}
f_{X,Y}(x, y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \\
&= \frac{1}{2\pi\sigma_X\sigma_Y} \cdot exp\left(-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \qquad (\rho = 0 \because X, Y \textit{ are uncorrelated}) \\
&= \frac{1}{\sqrt{2\pi}\sigma_X} \cdot exp\left(-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}\right]\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} \cdot exp\left(-\frac{1}{2}\left[\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right) \\
&= f_X(x) \cdot f_Y(y).
\end{aligned}$$

3. (3 pts) Give an example to show that two random variables X, Y can be marginally Gaussian (meaning, X is Gaussian, and Y is Gaussian) and uncorrelated but *not* independent. Hint: (X, Y) cannot be multivariate Gaussian in this case.

SOLUTION GOES HERE

Define $X \sim N(0, 1)$. Then define

$$Y = \begin{cases} X & \textit{with probability } \frac{1}{2} \\ -X & \textit{with probability } \frac{1}{2}. \end{cases}$$

We can verify that Y is Gaussian by checking its *CDF*

$$\begin{aligned} P(Y \le y) &= P(X \le y | Y = X)P(Y = X) + P(-X \le y | Y = -X)P(Y = -X) \\ &= P(X \le y) \cdot \frac{1}{2} + P(X \ge -y) \cdot \frac{1}{2} \qquad (\textit{notice } P(X \le y) = P(X \ge -y)) \\ &= P(X \le y). \end{aligned}$$

Hence, $Y \sim N(0, 1)$. Now we check that X, Y are uncorrelated by evaluating their covariance

$$\begin{aligned} Cov(X, Y) &= E[XY] - E[X]E[Y] \\ &= E[XY] - 0 \\ &= E[XY | Y = X]P(Y = X) + E[XY | Y = -X]P(Y = -X) \\ &= E[X^2] \cdot \frac{1}{2} + E[-X^2] \cdot \frac{1}{2} \\ &= 0. \end{aligned}$$

However, we know X, Y are not independent, because $P(Y \le -1 | -0.1 \le X \le 0.1) = 0 \ne P(Y \le -1)$.

Random walks

4. (4 pts) Let x_t, $t = 1, 2, 3, \ldots$ be a random walk with drift:

$$x_t = \delta + x_{t-1} + \epsilon_t,$$

where (say) $\epsilon_t \sim N(0, \sigma^2)$ for $t = 1, 2, 3, \ldots$. Suppose that both δ and σ^2 are unknown. Devise a test statistic for the null hypothesis that $\delta = 0$. This should be based on a standard test that you know (have learned in a past course) for testing whether the mean of Gaussian is zero, with unknown variance, based on i.i.d. samples from this Gaussian.

State what the null distribution is for this test statistic, and how you would compute it in R (a function name is sufficient if the test statistic is implemented as a function in base R). Hint: consider taking differences along the sequence ... after that, what you want sounds like "c-test", or "p-test", or "ϕ-test", or ...

SOLUTION GOES HERE

We will proceed by testing the difference of the random walk sequence. Let $d_t = x_t - x_{t-1} = \delta + \epsilon_t$. Then $d_t \sim N(0 + \delta, \sigma^2)$ and d_t is *i.i.d.* Since σ is unknown, we can use the *t-test* to test whether the mean of d_t is zero, which means $\delta = 0$. The null distribution is t_{n-1}, where n is the length of the random walk sequence. The test statistic is

$$t = \frac{\bar{d}}{\hat{\sigma}/\sqrt{n}}.$$

We can compute it in R by using the function `t.test()`.

5. (2 pts) Simulate a random walk of length 100 *without* drift, i.e., $\delta = 0$, and compute the test statistic you devised in Q4 and report its value. Then repeat, but using a large nonzero value δ.

```
# CODE GOES HERE
# Generate some data
n <- 101
delta <- 0
sigma <- 1
epsilon <- rnorm(n, 0, sigma)
x <- rep(0, n)
for (t in 2:n) {
```

```
  x[t] <- delta + x[t - 1] + epsilon[t]
}

# First differences
diff_x <- diff(x)

# t-test
test_result <- t.test(diff_x, mu = 0, alternative = "two.sided")
t_value <- test_result$statistic
p_value <- test_result$p.value

print(test_result)
```

```
##
##  One Sample t-test
##
## data:  diff_x
## t = 0.64714, df = 99, p-value = 0.519
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1394043  0.2743469
## sample estimates:
##  mean of x
## 0.06747132
```

```
print(paste("test statistic:", t_value))
```

```
## [1] "test statistic: 0.64714132140939"
```

```
print(paste("likelihood of null hypothesis", p_value))
```

```
## [1] "likelihood of null hypothesis 0.519038207910816"
```

```
# CODE GOES HERE
# Generate some data
# let drift be 10
n <- 101
delta <- 10
sigma <- 1
epsilon <- rnorm(n, 0, sigma)
x <- rep(0, n)
for (t in 2:n) {
  x[t] <- delta + x[t - 1] + epsilon[t]
}

# First differences
diff_x <- diff(x)

# t-test
test_result <- t.test(diff_x, mu = 0, alternative = "two.sided")
t_value <- test_result$statistic
p_value <- test_result$p.value

print(test_result)
```

```
##
```

```
##  One Sample t-test
##
## data:  diff_x
## t = 109.38, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.800815 10.162966
## sample estimates:
## mean of x
##   9.98189
```

```
print(paste("test statistic:", t_value))
```

```
## [1] "test statistic: 109.38111283583"
```

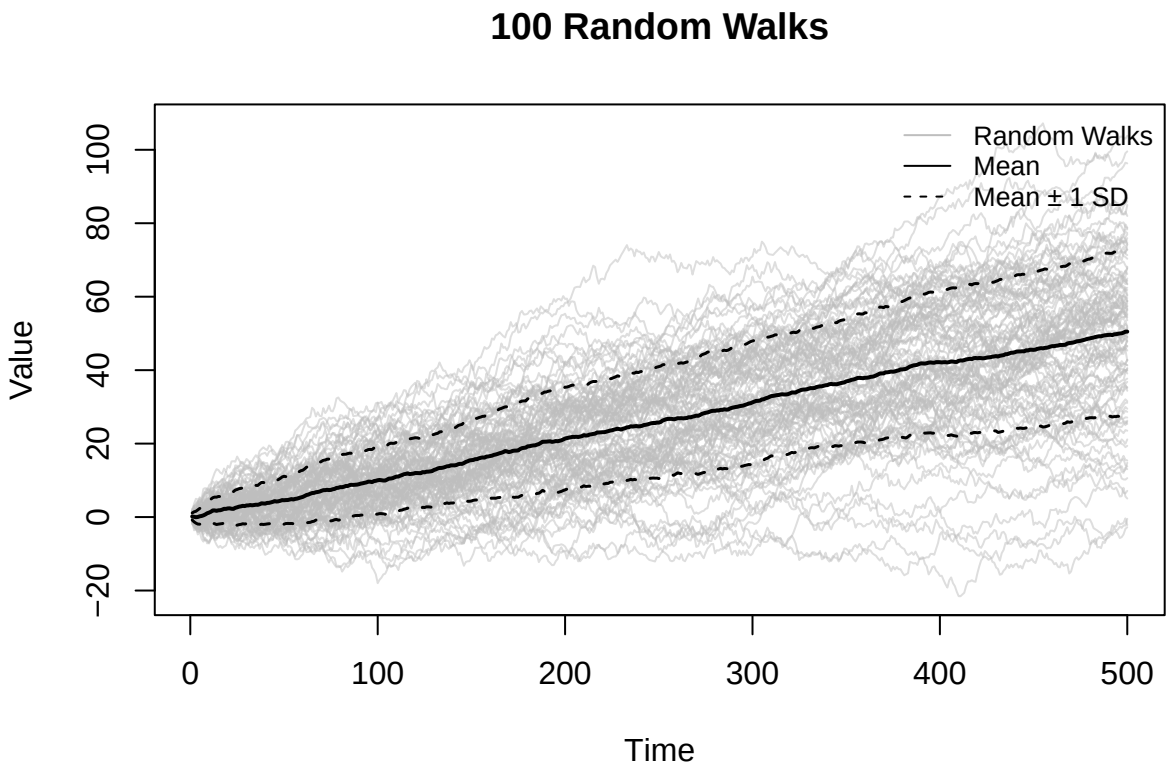
```
print(paste("likelihood of null hypothesis", p_value))
```

```
## [1] "likelihood of null hypothesis 4.53186535491104e-105"
```

6. (4 pts) Simulate 100 random walks each of length 500, with nonzero drift, and plot them on the same plot using transparent coloring, following the code used in the lecture notes from week 2 ("Measures of dependence and stationarity"). Calculate the sample mean $\hat{\mu}_t$ at each time t, across the repetitions, and plot as a dark line on the same plot. Then, calculate the sample standard deviation $\hat{\sigma}_t$ at each time t, and plot the mean plus or minus one standard deviation: $\hat{\mu}_t \pm \hat{\sigma}_t$, as dark dotted lines on the same plot. Describe what you see (you should see that both the mean and variance increase over time).

```
# CODE GOES HERE


n_simulations <- 100
n_steps <- 500
drift <- 0.1

# 1. Simulate 100 random walks each of length 500, with nonzero drift
random_walks <- matrix(0, n_simulations, n_steps)
for (i in 1:n_simulations){
  random_walks[i, ] <- cumsum(rnorm(n_steps, 0, 1) + drift)
}

# 2. Plot them on the same plot using transparent coloring
matplot(
    t(random_walks),
    type = "l",
    lty = 1,
    col = adjustcolor("grey", alpha.f = 0.5),
    main = "100 Random Walks",
    xlab = "Time",
    ylab = "Value"
)

# 3. Calculate the sample mean and add to the plot
sample_mean <- colMeans(random_walks)
lines(sample_mean, col = "black", lwd = 2)

# 4. Calculate the sample standard deviation and add to the plot
sample_sd <- apply(random_walks, 2, sd)
```

```r
lines(sample_mean + sample_sd, col = "black", lty = 2, lwd = 1.5)
lines(sample_mean - sample_sd, col = "black", lty = 2, lwd = 1.5)

legend("topright",
    legend = c("Random Walks", "Mean", "Mean ± 1 SD"),
    col = c("grey", "black", "black"),
    lty = c(1, 1, 2), cex = 0.8, bty = "n"
)
```



Stationarity

7. (3 pts) Compute the mean, variance, auto-covariance, and auto-correlation functions for the process

$$x_t = w_t w_{t-1},$$

where each $w_t \sim N(0, \sigma^2)$, independently. Is x_t, $t = 1, 2, 3, \ldots$ stationary?

SOLUTION GOES HERE

$$
\begin{aligned}
E[x_t] &= E[w_t w_{t-1}] \\
&= E[w_t]E[w_{t-1}] \qquad (\textit{because } w_t, w_{t-1} \textit{ are independent}) \\
&= 0 \cdot 0 \\
&= 0. \\
Var(X_t) &= Var(w_t w_{t-1}) \\
&= E[w_t^2 w_{t-1}^2] - E[w_t w_{t-1}]^2 \\
&= E[w_t^2]E[w_{t-1}^2] - 0^2 \qquad (\textit{because } w_t, w_{t-1} \textit{ are independent}) \\
&= \sigma^2 \cdot \sigma^2 \\
&= \sigma^4.
\end{aligned}
$$

For Auto-covariance, we will evaluate case by case for $h = 0, 1$ and else. For $h = 0$, we have

$$
\begin{aligned}
\gamma_x(0) &= Cov(x_t, x_t) \\
&= Var(x_t) \\
&= \sigma^4.
\end{aligned}
$$

For $h = 1$, we have

$$
\begin{aligned}
\gamma_x(1) &= Cov(x_t, x_{t+1}) \\
&= E[w_t w_{t-1} w_{t+1} w_t] - E[x_t]E[x_{t+1}] \\
&= E[w_t^2]E[w_{t-1}]E[w_{t+1}] - 0 \cdot 0 \qquad (\textit{because } w_t, w_{t-1}, w_{t+1} \textit{ are independent}) \\
&= \sigma^2 \cdot 0 \cdot 0 \\
&= 0.
\end{aligned}
$$

We can easily tell that for $h < -1, h > 1$, x_t, x_{t+h} are completely independent. Hence,

$$
\gamma_x(h) = \begin{cases} \sigma^4 & \textit{if } h = 0 \\ 0 & \textit{else.} \end{cases}
$$

Normalize the auto-covariance with the variance, we have

$$
\rho_x(h) = \begin{cases} \frac{\sigma^4}{\sigma^4} = 1 & \textit{if } h = 0 \\ 0 & \textit{else.} \end{cases}
$$

Since $\mu_{x,y} = 0$ is constant and $\gamma_x(x_t, x_{t+h})$ is only dependent on h, we can conclude that x_t is stationary.

8. (3 pts) Repeat the same calculations in Q7, but where each $w_t \sim N(\mu, \sigma^2)$, independently, for $\mu \neq 0$. Is x_t, $t = 1, 2, 3, \ldots$ stationary?

SOLUTION GOES HERE

$$
\begin{aligned}
E[x_t] &= E[w_t w_{t-1}] \\
&= E[w_t]E[w_{t-1}] \qquad (\textit{because } w_t, w_{t-1} \textit{ are independent}) \\
&= \mu \cdot \mu \\
&= \mu^2. \\
Var(X_t) &= Var(w_t w_{t-1}) \\
&= E[w_t^2 w_{t-1}^2] - E[x_t]^2 \\
&= E[w_t^2]E[w_{t-1}^2] - \mu^4 \qquad (\textit{because } w_t, w_{t-1} \textit{ are independent}) \\
&= (\sigma^2 + \mu^2)^2 - \mu^4 \\
&= \sigma^4 + 2\sigma^2\mu^2.
\end{aligned}
$$

For Auto-covariance, we will evaluate case by case for $h = 0, 1$ and else. For $h = 0$, we have

$$
\begin{aligned}
\gamma_x(0) &= Cov(x_t, x_t) \\
&= Var(x_t) \\
&= \sigma^4 + 2\sigma^2\mu^2.
\end{aligned}
$$

For $h = 1$, we have

$$
\begin{aligned}
\gamma_x(1) &= Cov(x_t, x_{t+1}) \\
&= E[w_t w_{t-1} w_{t+1} w_t] - E[x_t]E[x_{t+1}] \\
&= E[w_t^2]E[w_{t-1}]E[w_{t+1}] - \mu^2 \cdot \mu^2 \qquad (\textit{because } w_t, w_{t-1}, w_{t+1} \textit{ are independent}) \\
&= (\sigma^2 + \mu^2)(\mu)(\mu) - \mu^4 \\
&= \sigma^2\mu^2.
\end{aligned}
$$

We can easily tell that for $h < -1, h > 1$, x_t, x_{t+h} are completely independent. Hence,

$$
\gamma_x(h) = \begin{cases} \sigma^4 + 2\sigma^2\mu^2 & \textit{if } h = 0 \\ \sigma^2\mu^2 & \textit{if } h = 1, -1 \\ 0 & \textit{else.} \end{cases}
$$

Normalize the auto-covariance with the variance, we have

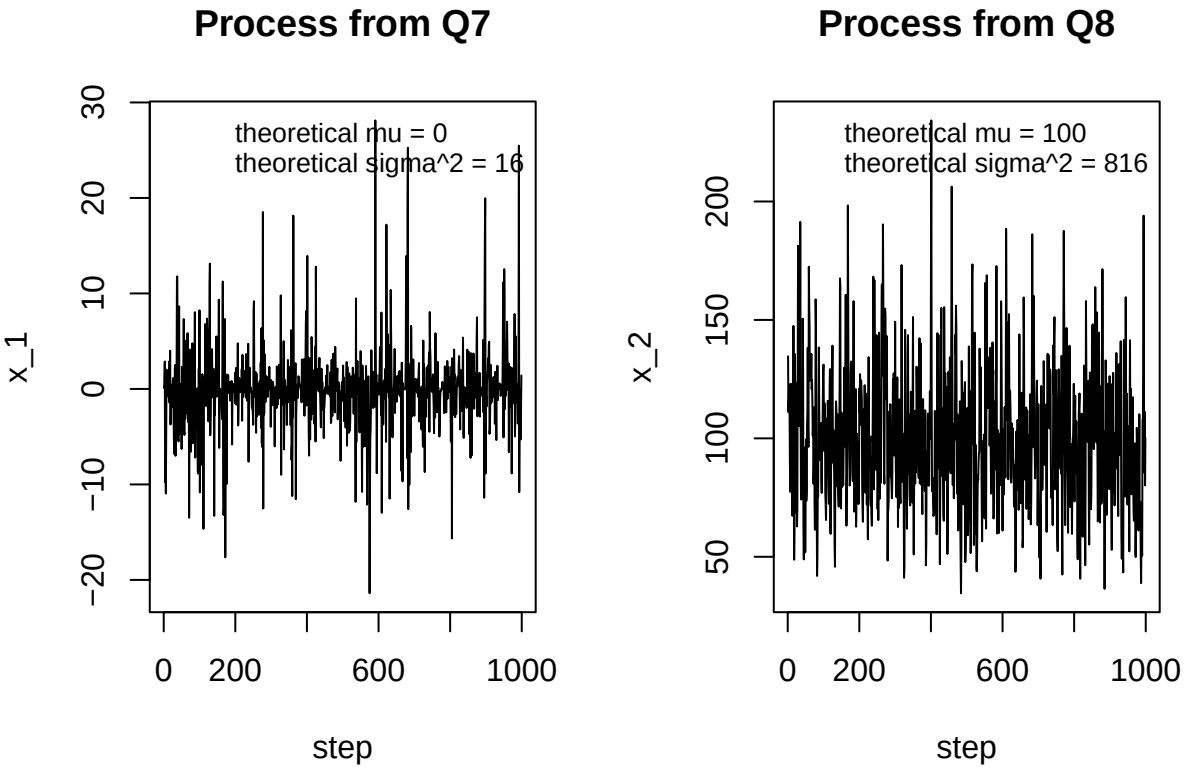
$$
\rho_x(h) = \begin{cases} 1 & \textit{if } h = 0 \\ \frac{\sigma^2\mu^2}{\sigma^4 + 2\sigma^2\mu^2} & \textit{else.} \end{cases}
$$

Since $\mu_{x,y} = \mu^2$ is constant and $\gamma_x(x_t, x_{t+h})$ is only dependent on h, we can conclude that x_t is stationary.

9. (3 pts) Simulate the processes from Q7 (with $\mu = 0$) and Q8 (with $\mu \neq 0$) and plot the results. Compute the sample mean and sample variance for each simulated process and check that it is close to the population mean and variance from Q7 and Q8. Compute and plot the sample auto-correlation function using `acf()` and check again that it agrees with the population auto-correlation function from Q7 and Q8.

```
# CODE GOES HERE
n_steps <- 1000
mu <- 10
sigma <- 2
norm_0 <- rnorm(n_steps, 0, sigma)
```

```r
norm_mu <- rnorm(n_steps, mu, sigma)

x_1 <- norm_0[-1] * norm_0[-n_steps]
x_2 <- norm_mu[-1] * norm_mu[-n_steps]

par(mfrow = c(1, 2))
plot(x_1, type = "l", main = "Process from Q7", xlab = "step")
legend("topright",
    legend = c("theoretical mu = 0", "theoretical sigma^2 = 16"),
    cex = 0.8, bty = "n"
)
plot(x_2, type = "l", main = "Process from Q8", xlab = "step")
legend("topright",
    legend = c("theoretical mu = 100", "theoretical sigma^2 = 816"),
    cex = 0.8, bty = "n"
)
```




```r
cat("Sample mean of Process 1 (Q7):", mean(x_1), "\n")
```

```
## Sample mean of Process 1 (Q7): -0.1375469
```

```r
cat("Sample variance of Process 1 (Q7):", var(x_1), "\n")
```

```
## Sample variance of Process 1 (Q7): 17.56031
```

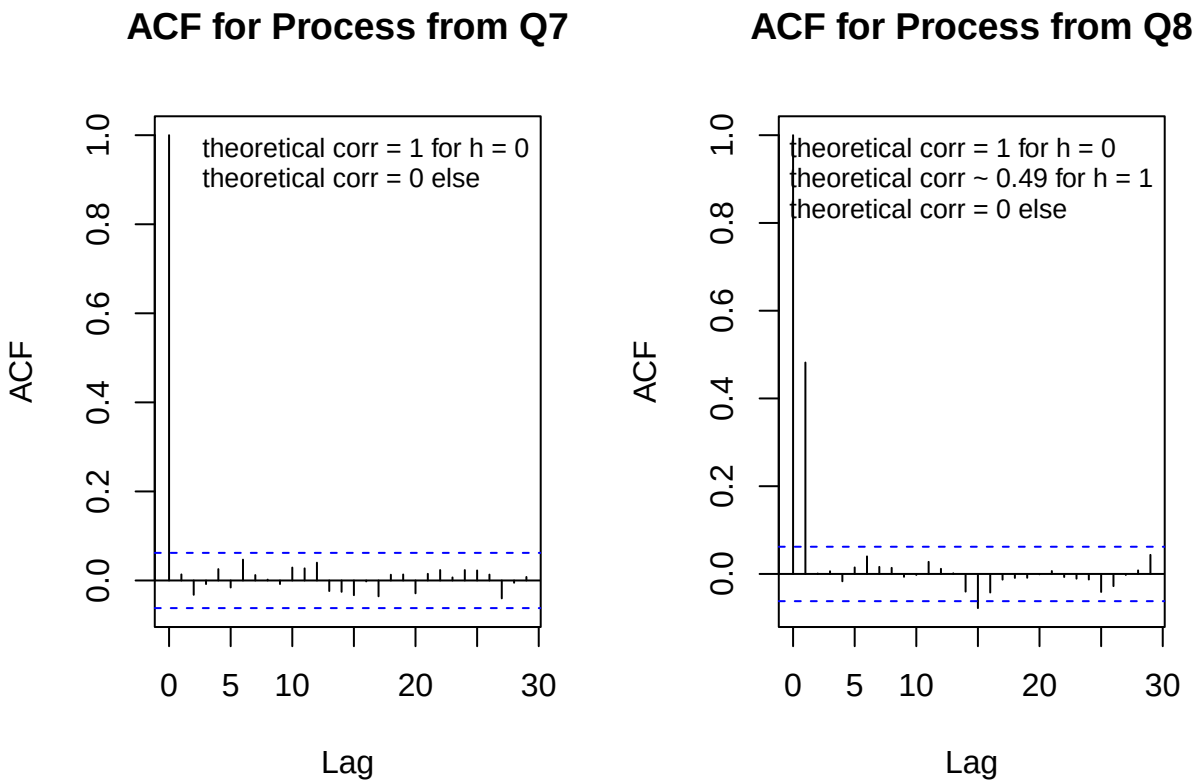
```r
cat("Sample mean of Process 2 (Q8):", mean(x_2), "\n")
```

```
## Sample mean of Process 2 (Q8): 99.39516
```

```r
cat("Sample variance of Process 2 (Q8):", var(x_2), "\n")
```

```
## Sample variance of Process 2 (Q8): 820.3406
```

```r
par(mfrow = c(1, 2))
acf(x_1, main = "ACF for Process from Q7")
legend("topright",
    legend = c("theoretical corr = 1 for h = 0", "theoretical corr = 0 else"),
    cex = 0.8, bty = "n"
)
acf(x_2, main = "ACF for Process from Q8")
legend("topright",
    legend = c("theoretical corr = 1 for h = 0",
    "theoretical corr ~ 0.49 for h = 1", "theoretical corr = 0 else"),
    cex = 0.8, bty = "n"
)
```



10. (2 pts) Give an example of a weakly stationary process that is not strongly stationary.

SOLUTION GOES HERE

Define a time series x_t such that when t is odd, $x_t \sim N(0,1)$, and when t is even, $x_t \sim Uniform([-\sqrt{3}, \sqrt{3}])$, and x_t is independent for all t. Then, x_t is weakly stationary because its mean and variance are constant. Also, the covariance with lag $h \neq 0$ is 0 because x_t are independent. However, x_t is not strongly stationary because obviously $x_t \not\sim x_{t+1}$

11. (Bonus) A function κ is said to be *positive semidefinite* (PSD) provided that

$$\sum_{i,j=1}^{n} a_i a_j \kappa(t_i - t_j) \geq 0, \quad \text{for all } n \geq 1, \text{ all } a_1, \ldots, a_n, \text{ and all } t_1, \ldots, t_n.$$

Prove that if x_t, $t = 1, 2, 3, \ldots$ is stationary, and $\gamma_x(h)$ is its auto-covariance function (as a function of lag h), then γ_x is PSD. You may use any equivalences between PSD and other linear-algebraic properties that you want, as long as you state clearly what you are using.

SOLUTION GOES HERE

We will take the variance of the linear combination of the time series, and we will realize that it is in the required form.

$$\begin{aligned} Var(\sum_{i=1}^{n} a_i \cdot x_{t_i}) &= \sum_{i=1}^{n} a_i^2 \cdot Var(x_{t_i}) + \sum_{i \neq j} a_i a_j \cdot Cov(x_{t_i}, x_{t_j}) \\ &= \sum_{i=1}^{n} a_i^2 \cdot Cov(x_{t_i}, x_{t_i}) + \sum_{i \neq j} a_i a_j \cdot Cov(x_{t_i}, x_{t_j}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \cdot Cov(x_{t_i}, x_{t_j}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \cdot \gamma_x(t_i - t_j). \end{aligned}$$

Notice that the variance is always greater than or equal to 0. Hence, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \cdot \gamma_x(t_i - t_j) \geq 0.$$

Indeed, γ_x is PSD.

Side note: the reason it works here is because $Cov(x_{t_i}, x_{t_j})$ can be written as a function of $t_i - t_j$ under the stationary assumption. If it is not the case, we cannot use this method to prove PSD.

Joint stationarity

Notions of joint stationarity, between two time series, can be defined in an analogous way to how we defined stationarity in lecture. We say that two time series x_t, $t = 1, 2, 3, \ldots$ and y_t, $t = 1, 2, 3, \ldots$ are *strongly jointly stationary* provided that:

$$(x_{s_1}, x_{s_2}, \ldots, x_{s_k}, y_{t_1}, y_{t_2}, \ldots, y_{t_\ell}) \stackrel{d}{=} (x_{s_1+h}, x_{s_2+h}, \ldots, x_{s_k+h}, y_{t_1+h}, y_{t_2+h}, \ldots, y_{t_\ell+h}),$$
$$\text{for all } k, \ell \geq 1, \text{ all } s_1, \ldots, s_k \text{ and } t_1, \ldots, t_\ell, \text{ and all } h.$$

Here $\stackrel{d}{=}$ means equality in distribution. In other words, any collection of variates from the two sequences has the same joint distribution after we shift the time indices forward or backwards in time. Meanwhile, we say that x_t, $t = 1, 2, 3, \ldots$ and y_t, $t = 1, 2, 3, \ldots$ are *weakly jointly stationary* or simply *jointly stationary* provided that each series is stationary, and:

$$\gamma_{xy}(s, t) = \gamma_{xy}(s + h, t + h), \quad \text{for all } s, t, h.$$

Here γ_{xy} is the cross-covariance function between x, y. In other words, the cross-covariance function must be invariant to shifts forward or backwards in time, and is only a function of the lag $h = s - t$. For jointly stationary series, we can hence abbreviate their cross-covariance function by $\gamma_{xy}(h)$.

SOLUTION GOES HERE

12. (2 pts) Give an example of two time series that are weakly jointly stationary but not strongly jointly stationary.

SOLUTION GOES HERE

Define a time series x_t the same as in Q10, and define $y_{s+1} = x_s$. Then, as proved in Q10, both x and y are weekly stationary. Also, the cross covariance $\gamma_{xy}(s,t) = 1$ for $t = s+1$ and $\gamma_{xy}(s,t) = 0$ for $t \neq s+1$ because of their independence. Hence, $\gamma_{xy}(s,t)$ is only dependent on $s-t$ and is invariant to any shift in time. However, x and y are not strongly jointly stationary. A counter-example would be to take $(x_t, y_{t-1}) \not\sim (x_{t+1}, y_t)$ because one pair would be a joint independent gaussian distribution and the other would be a joint independent uniform distribution.

13. (3 pts) If x_t, $t = 1, 2, 3, \ldots$ and y_t, $t = 1, 2, 3, \ldots$ form a *joint Gaussian process*, which means that any collection $(x_{s_1}, x_{s_2}, \ldots, x_{s_k}, y_{t_1}, y_{t_2}, \ldots, y_{t_\ell})$ of variates along the series has a multivariate Gaussian distribution, then prove that weak joint stationarity implies strong joint stationarity.

SOLUTION GOES HERE

Notice that a multivariate Gaussian distribution is completely determined by its mean vector and covariance matrix. If the mean vector and covariance matrix are invariant to time shift, then the distribution is invariant to time shift, hence the process is strongly jointly stationary.

We will first prove that the mean vector is invariant to time shift. Since the times series are weakly jointly stationary, they are marginally weakly stationary. Hence, μ_x and μ_y are constant. Now we evaluate the mean vector of $(x_{s_1}, x_{s_2}, \ldots, x_{s_k}, y_{t_1}, y_{t_2}, \ldots, y_{t_\ell})$. Indeed, the mean of each coordinate is μ_x or μ_y, which are constant. Hence, the mean vector is invariant to time shift.

We will then prove that the covariance matrix is invariant to time shift. Notice the entries of the covariance matrix is determined by the covariance between each coordinate in the $(x_{s_1}, x_{s_2}, \ldots, x_{s_k}, y_{t_1}, y_{t_2}, \ldots, y_{t_\ell})$. If the covariance is between x, the covariance is invariant of time shift because x is weakly stationary, and the same holds for y. If the covariance is between x, y, the covariance is also invariant to time shift, because x, y are weakly jointly stationary. Hence, covariance between any two arbitrary coordinates are invariant to time shift. Hence, the covariance matrix is invariant to time shift.

14. (3 pts) Write down explicit formulas that shows how to estimate the cross-covariance and cross-correlation function of two finite time series $x_t, t = 1, \ldots, n$ and y_t, $t = 1, \ldots, n$, under the assumption of joint stationarity. Hint: these should be analogous to the *sample auto-covariance and sample auto-correlation functions* that we covered in lecture.

SOLUTION GOES HERE

Define $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ to be the sample mean of x and $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ to be the sample mean of y.

For positive lag h such that y is ahead of x or $h = 0$ such that x and y are at the same time, we have:

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x})(y_{t+h} - \bar{y}).$$

For negative lag h such that y is behind x, we have:

$$\hat{\gamma}_{xy}(h) = \frac{1}{n}\sum_{t=1-h}^{n}(x_t - \bar{x})(y_{t+h} - \bar{y}).$$

For correlation, we have:

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\hat{\sigma}_x \hat{\sigma}_y}$$

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$

$$\hat{\sigma}_y^2 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2.$$

15. (4 pts) Following the code used in the lecture notes from week 2 ("Measures of dependence and stationarity"), use the `ccf()` function to compute and plot the sample cross-correlation function between Covid-19 cases and deaths, separately, for each of Florida, Georgia, New York, Pennsylvania, and Texas. (The lecture code does this for California.) Comment on what you find: do the cross-correlation patterns look similar across different states?

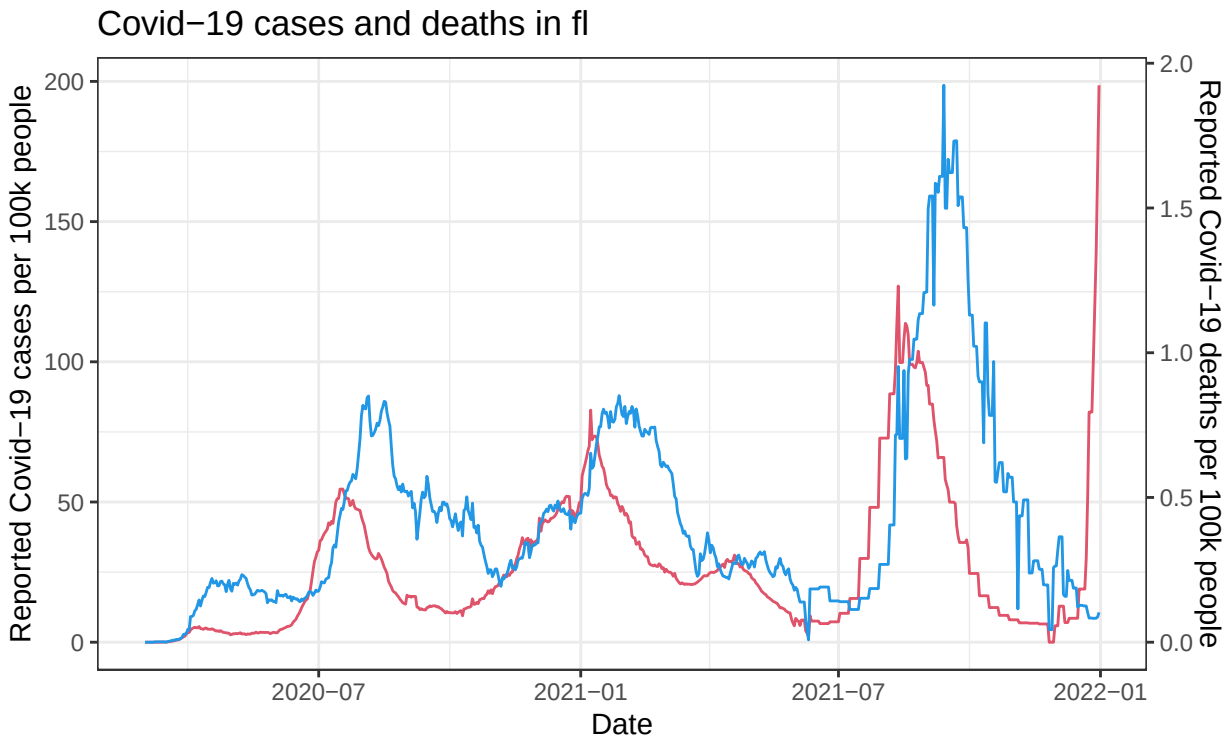
    Also, follow the lecture code to plot the case and death signals together, on the same plot, for each state (the lecture code provides a way to do this so that they are scaled dynamically to attain the same min and max, and hence look nice when plotted together). Comment on whether the estimated cross-correlation patterns agree with what you see visually between the case and death signals.

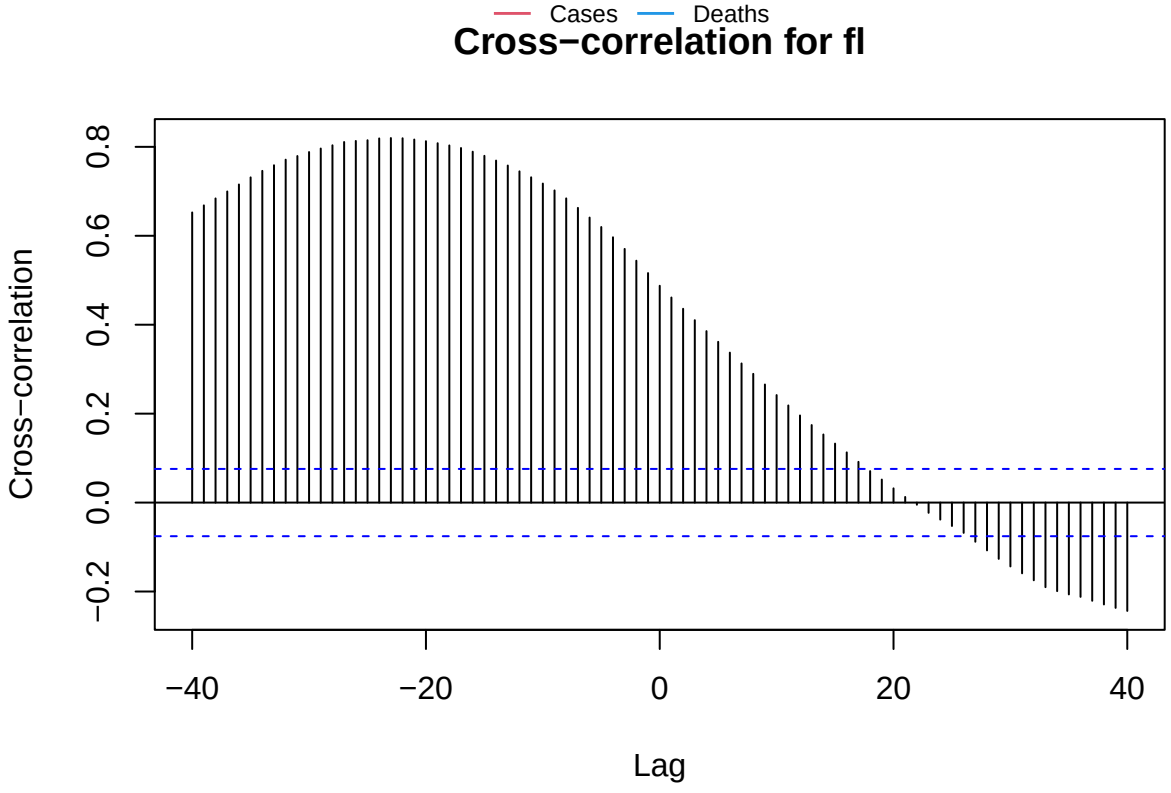
```
# CODE GOES HERE
library(epidatasets)
library(tidyverse)
```

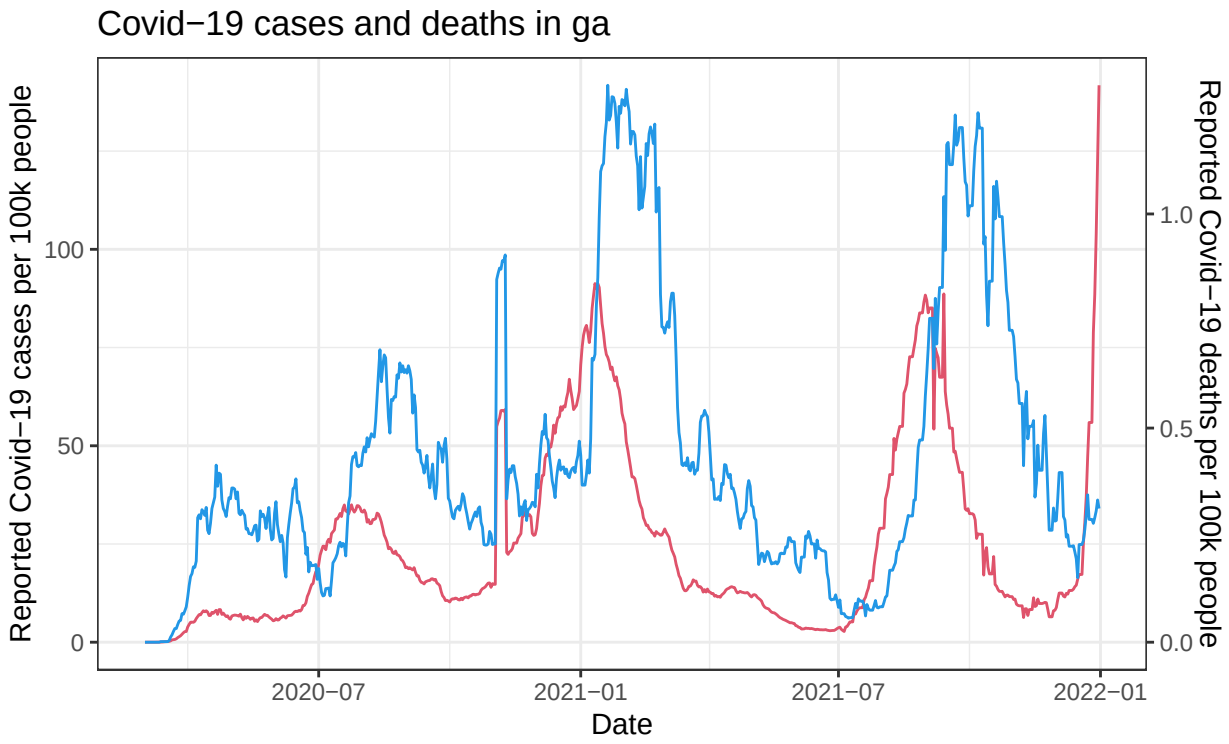
```
## -- Attaching core tidyverse packages ------------------------------------------------------------
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

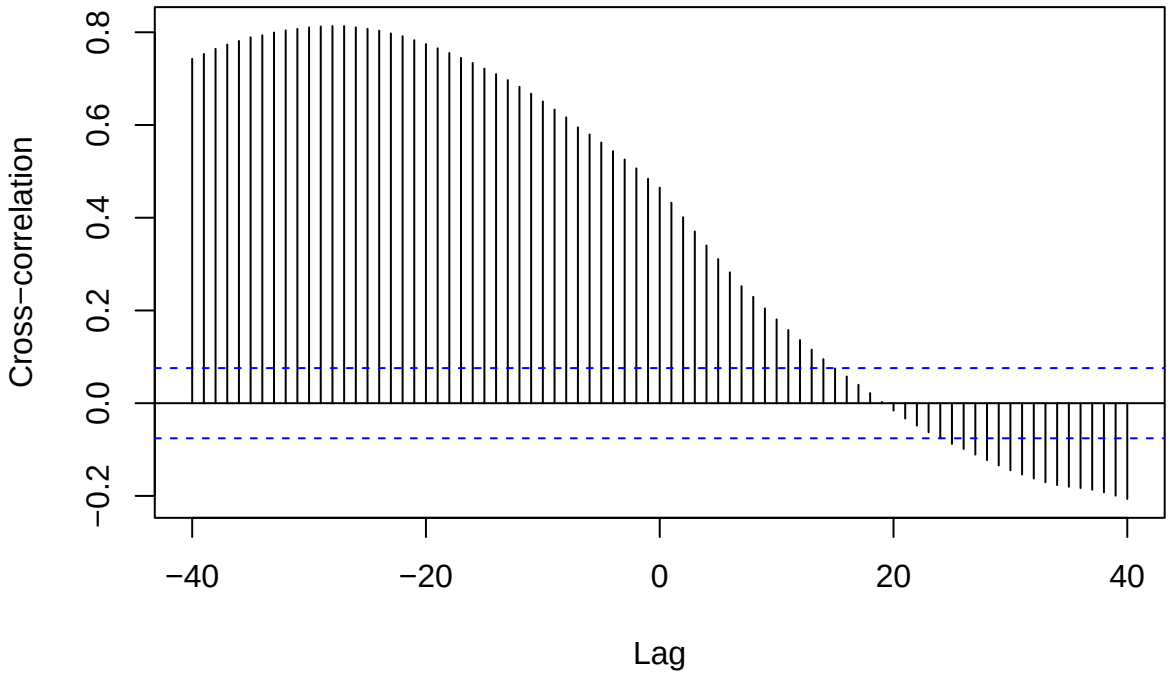
```
plot_state_data <- function(state) {
    # Covid-19 cases and deaths in California, pivot longer
    df = cases_deaths_subset |>
        filter(geo_value == state) |>
        select(time_value, case_rate_7d_av, death_rate_7d_av) |>
        pivot_longer(cols = c(case_rate_7d_av, death_rate_7d_av)) |>
        mutate(name = recode(name,
                             case_rate_7d_av = "Cases",
                             death_rate_7d_av = "Deaths"))

    # Handy function to produce a transformation from one range to another
    trans = function(x, from_range, to_range) {
        (x - from_range[1]) / (from_range[2] - from_range[1]) *
            (to_range[2] - to_range[1]) + to_range[1]
    }

    # Compute ranges of the two signals, and transformations in b/w them
    range1 = df |> filter(name == "Cases") |> select("value") |> range()
    range2 = df |> filter(name == "Deaths") |> select("value") |> range()
    trans12 = function(x) trans(x, range1, range2)
    trans21 = function(x) trans(x, range2, range1)

    ggplot_object <- ggplot(
        bind_rows(
            df |> filter(name == "Cases"),
            df |> filter(name == "Deaths") |> mutate_at("value", trans21)
        ),
        aes(x = time_value, y = value)
    ) +
        geom_line(aes(color = name)) +
```

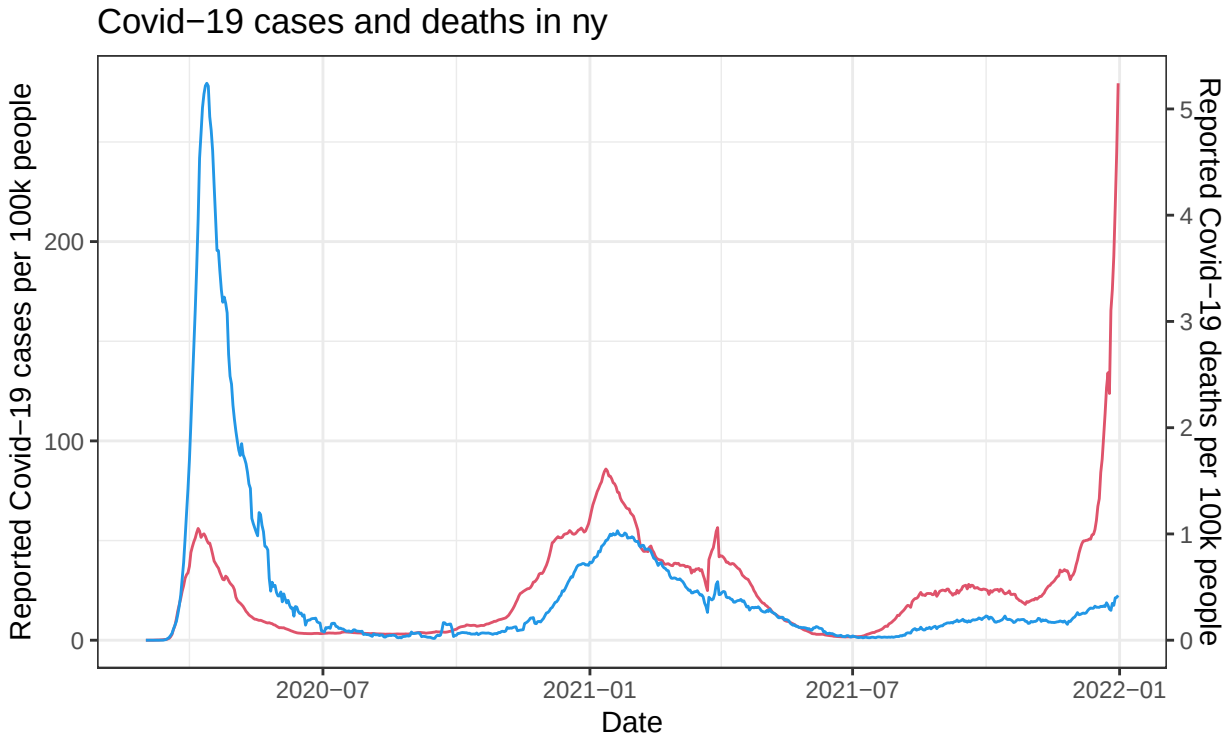
```
        scale_color_manual(values = palette()[c(2, 4)]) +
        scale_y_continuous(
            name = "Reported Covid-19 cases per 100k people",
            limits = range1,
            sec.axis = sec_axis(
                trans = trans12,
                name = "Reported Covid-19 deaths per 100k people"
            )
        ) +
        labs(
            title = paste("Covid-19 cases and deaths in", state),
            x = "Date"
        ) +
        theme_bw() +
        theme(legend.position = "bottom", legend.title = element_blank())

    print(ggplot_object)

    return(df)
}

states <- c("fl", "ga", "ny", "pa", "tx")
for (state in states) {
  df <- plot_state_data(state)
  ccf(df |> filter(name == "Cases") |> select(value),
      df |> filter(name == "Deaths") |> select(value),
      lag.max = 40, ylab = "Cross-correlation", main = ""
  )
  title(main = paste("Cross-correlation for", state))
}
```

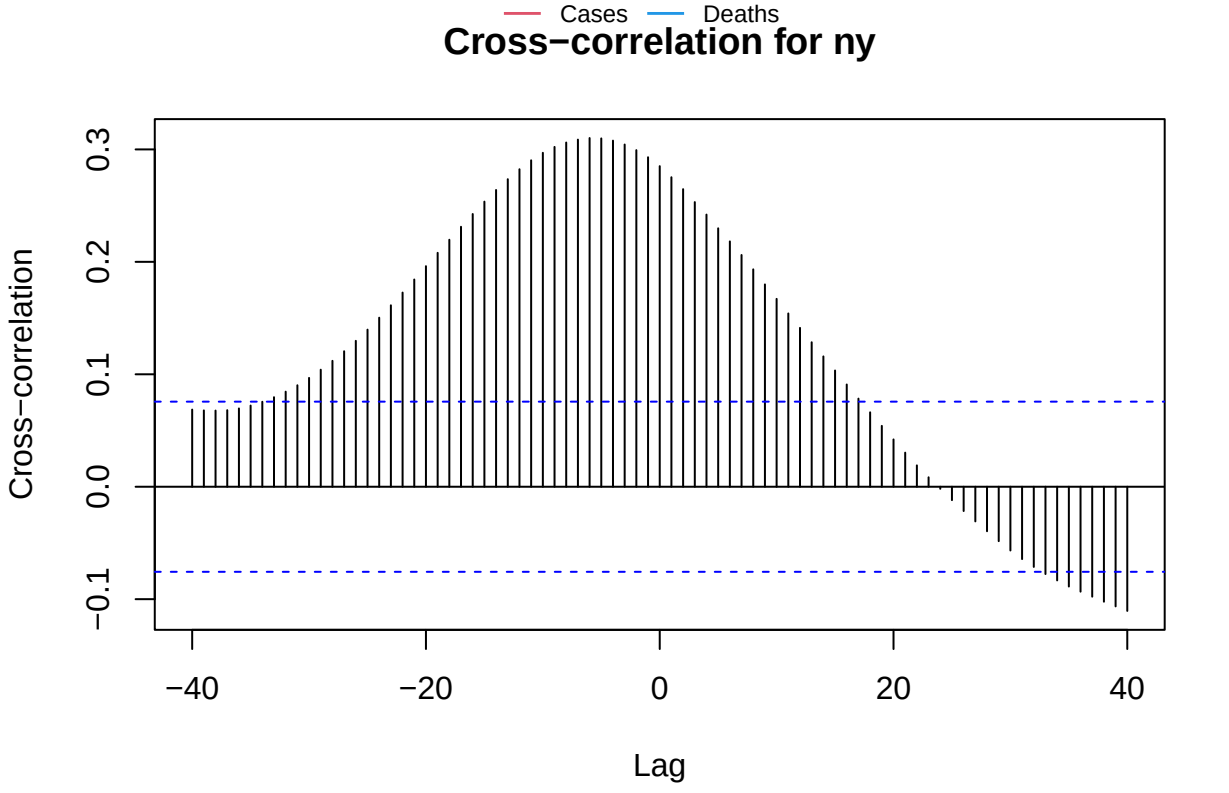

Covid−19 cases and deaths in fl
Reported Covid−19 cases per 100k people
Reported Covid−19 deaths per 100k people
200
150
100
50
0
2.0
1.5
1.0
0.5
0.0
2020−07
2021−01
2021−07
2022−01
Date
Cases
Deaths

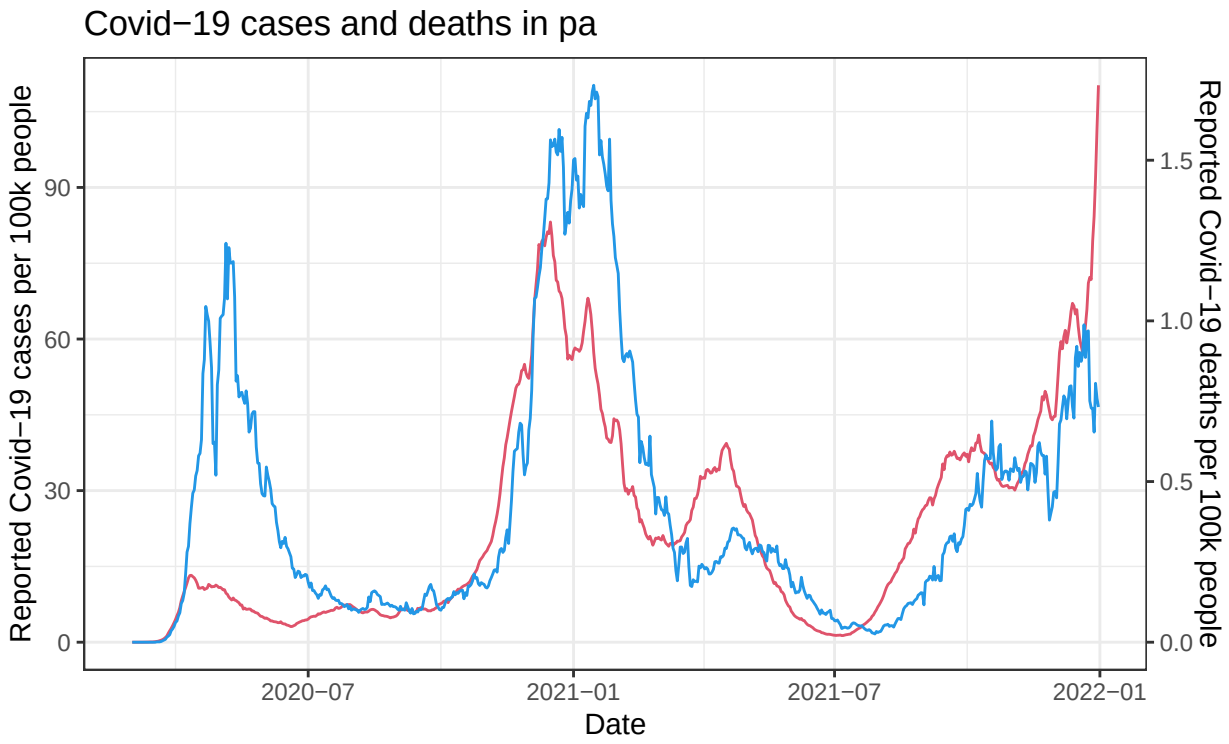

Cross−correlation for fl
Cross−correlation
0.8
0.6
0.4
0.2
0.0
−0.2
−40
−20
0
20
40
Lag

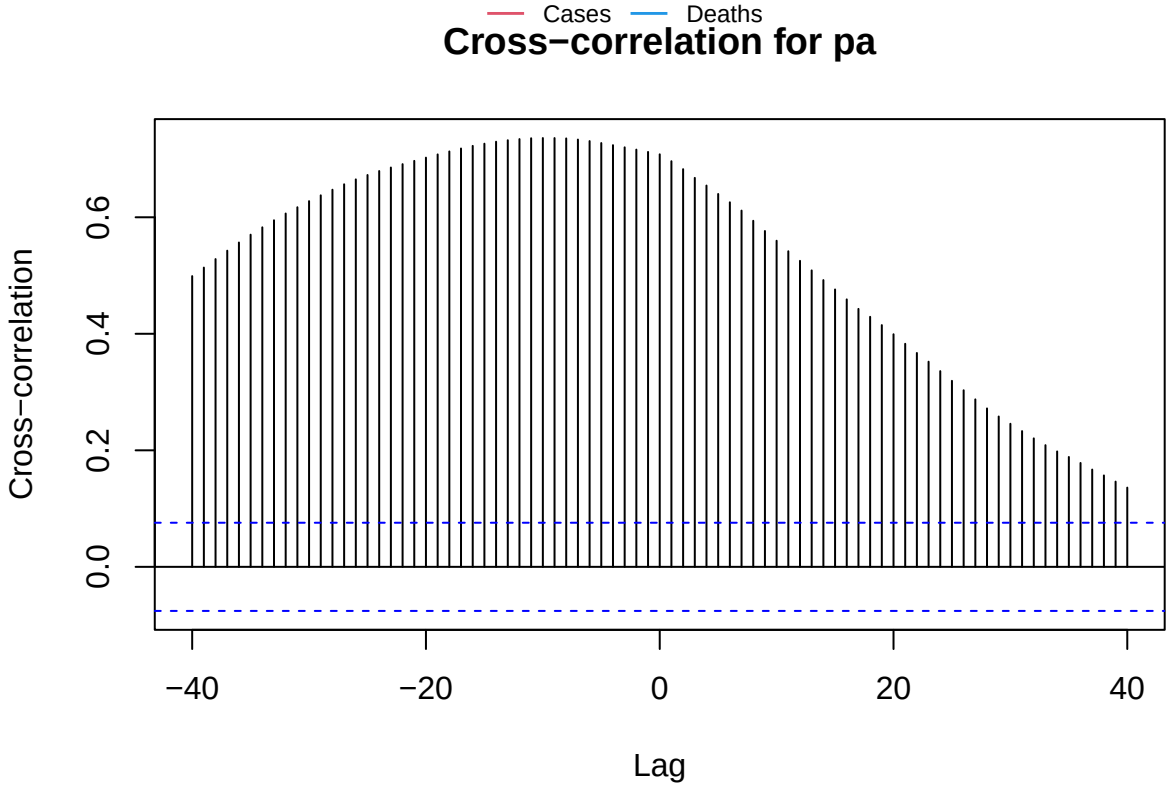

Covid−19 cases and deaths in ga
Reported Covid−19 cases per 100k people
Reported Covid−19 deaths per 100k people
0
50
100
0.0
0.5
1.0
2020−07
2021−01
2021−07
2022−01
Date
Cases
Deaths

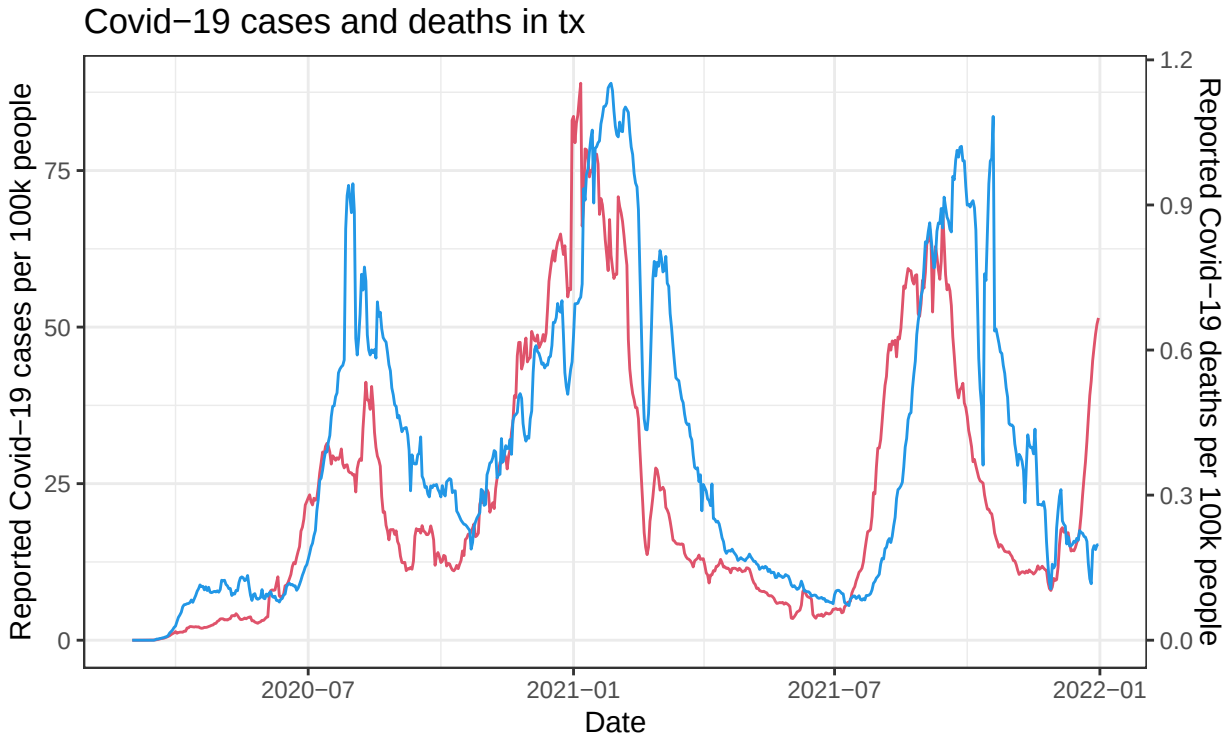

Cross−correlation for ga
Cross−correlation
−0.2
0.0
0.2
0.4
0.6
0.8
−40
−20
0
20
40
Lag

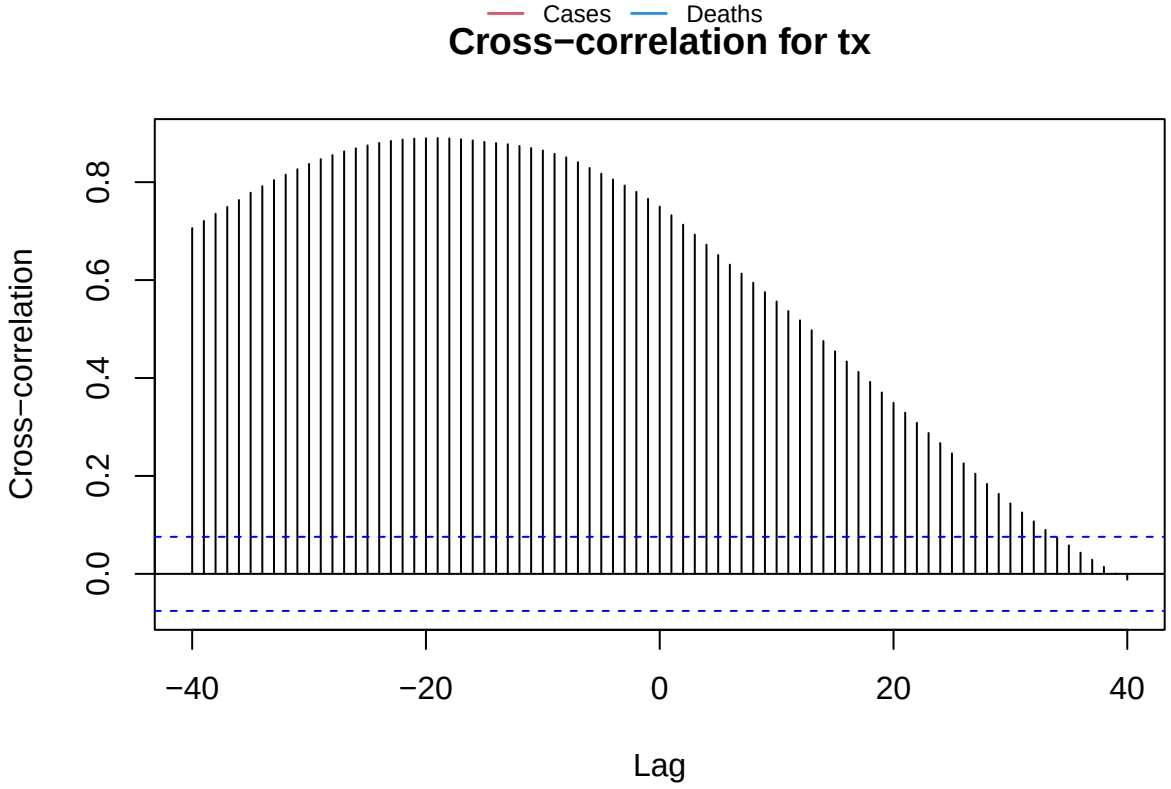

Covid−19 cases and deaths in ny
Reported Covid−19 cases per 100k people
Reported Covid−19 deaths per 100k people
0
100
200
0
1
2
3
4
5
2020−07
2021−01
2021−07
2022−01
Date
Cases
Deaths


Cross−correlation for ny
Cross−correlation
−0.1
0.0
0.1
0.2
0.3
−40
−20
0
20
40
Lag


Covid−19 cases and deaths in pa
Reported Covid−19 cases per 100k people
Reported Covid−19 deaths per 100k people
0
30
60
90
0.0
0.5
1.0
1.5
2020−07
2021−01
2021−07
2022−01
Date
Cases
Deaths


Cross−correlation for pa
Cross−correlation
0.0
0.2
0.4
0.6
−40
−20
0
20
40
Lag





Across different states, the cross-correlation patterns do look similar. This is unsurprising as lag between the death after case report should generally be consistent. We are expecting that death (if any) should occur some time after the case is reported and never lead the case.

The cross-correlation patterns do agree with what we can see visually between the case and death signals. For

some states, such as Florida, Georgia, and Texas, we can see the death signals generally lag behind the case signals. For those states, we can see that the cross-correlation peaked at around -20 days. Meanwhile, for the other states that the death signals are closer to the case signals, the cross-correlation peaked at around -5 day.