Homework 1: [YOUR NAME HERE]

Introduction to Time Series, Fall 2024

Due Friday September 13

The total number of points possible for this homework is 42. The number of points for each question is written below, and questions marked as "bonus" are optional (points awarded for bonus problems can be used to earn back points that you may have lost on other parts of this homework but will not put you above full credit). Submit the **knitted pdf 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).

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

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.

Random walks

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

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

where (say) $w_t \sim N(0, \sigma^2)$ for $t = 1, 2, 3, \ldots$. Recall from lecture that this is not stationary. Prove that $\rho(t-1, t) = \sqrt{\frac{t-1}{t}}$. What does this approach as $t \to \infty$ and what is the interpretation of this result?

5. (3 pts) Suppose that both δ and σ^2 are unknown. Devise a test statistic for the null hypothesis that $\delta = 0$ in the random walk model from Q4. 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 …

6. (2 pts) Simulate a random walk of length 200 *without* drift, i.e., $\delta = 0$, and compute the test statistic you devised in Q5 and report its value. Then repeat, but using a large nonzero value δ.

7. (4 pts) Simulate 50 random walks each of length 200, 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).

Stationarity

8. (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?

9. (3 pts) Repeat the same calculations in Q8, but where each $w_t \sim N(\mu, \sigma^2)$, independently, for $\mu \neq 0$. Is x_t, $t = 1, 2, 3, \ldots$ stationary?

10. (3 pts) Simulate the processes from Q8 (with $\mu = 0$) and Q9 (with $\mu \neq 0$), yielding two time series of length 200, and plot the results. Compute the sample mean and sample variance for each one (to be clear, this is just a sample mean of all data, over all time, and similarly for the variance), and check that these are close to the population mean and variance from Q8 and Q9. Also compute and plot the sample auto-correlation function using `acf()`, and check again that it agrees with the population auto-correlation function from Q8 and Q9.

11. (2 pts) Give an example of a weakly stationary process that is not strongly stationary.

12. (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 whatever elementary probability and/or linear algebra facts that you would like, as long as you state clearly what you are using.

13. (Bonus) Prove moreover that the sample auto-covariance function $\hat{\gamma}_x$ defined in lecture is also 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 $\gamma_{xy}(h)$.

14. (2 pts) Give an example of two time series that are weakly jointly stationary but not strongly jointly stationary.

15. (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.

16. (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 entirely analogous to the *sample auto-covariance and sample auto-correlation functions* that we covered in lecture.

17. (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.