
STAT/MATH 4/5540 - Project 3 - Spring 2020

by: Manuel Lladser

INSTRUCTIONS. Failure to follow these instructions may result in points discounted.

This project is due on Saturday, May 2-nd at 7 PM.

Discussing this project with anyone besides your group partner (if any), the Instructor, or the TA is not permitted. By submitting a report, all its participants agree to comply with the CU Honor Code Policy.

Students registered for APPM 4540 may work in groups of up to 2 members, and submit one project report with all participant names on it. Submit a single report in CANVAS. **Due to the COVID-19 (coronavirus) pandemic, avoid meeting in person and instead collaborate remotely using some video conferencing such as Facetime, Skype, WhatsApp, or Zoom.**

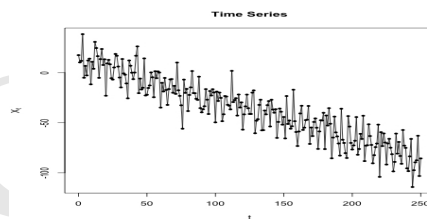
Students registered for APPM 5540 must work on the project on their own.

Your report is limited to 5 pages with a minimum font size of 11 points and 1-inch margins; in particular, please provide complete but brief answers. Appendices do not count toward the page limit!

To receive full credit, you must submit a professional report addressing all the instructions and questions in the same order as requested. Be sure to include all figures or tables and to label them (e.g., Figure 1, Table 2, etc). Also, be sure to cite any sources (textbooks, papers, websites, etc) you consult. Your write-up should include brief but complete answers to all the questions listed below with appropriate references to labeled figures or tables. At the end of your report, include an appendix with the R code you used to address the project. The code must include annotations. Good luck!

—O—

In this final project you will partially analyze the following time-series (download the file PRO3DATA.TXT in Canvas for the dataset):



The data was generated as follows:

$$X_t = a + b \cdot t + \sum_{j=1}^k c_j \cdot \cos(\omega_j t) + Z_t, \text{ for } t = 0, \dots, 250;$$

where a , b , k , the c_j 's, and the ω_j 's are constants, and $\{Z_t\} \sim \text{IDD}(0, \sigma^2)$, with $\sigma^2 > 0$. *All these constants are fixed but unreported to you!*

It is assumed that $0 < \omega_j < \pi$ for $j = 1, \dots, k$. **The goal of the project is to estimate these frequencies as well as their total number.** (Once we have estimates for these, the other constants may be estimated using standard regression techniques.) For this, it would be ideal if we could get rid of the noise $\{Z_t\}$. Unfortunately, this is not possible! However, we may use a *moving average filter* to damp down the noise. So let $0 < \ell < 125$, and consider:

$$X'_t := \frac{1}{2\ell + 1} \sum_{j=-\ell}^{\ell} X_{t+j}, \text{ for } t = \ell, \dots, (250 - \ell).$$

1. Using that $\cos(\alpha \pm \beta) = \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta)$, show that:

$$X'_t = a + b \cdot t + \sum_{j=1}^k (u'_j \cdot \cos(\omega_j t) + v'_j \cdot \sin(\omega_j t)) + Z'_t,$$

for suitable constants u'_j and v'_j , and a weakly stationary mean-zero time series $\{Z'_t\}$.

2. Explain why for ℓ large enough:

$$X'_t \approx a + b \cdot t + \sum_{j=1}^k (u'_j \cdot \cos(\omega_j t) + v'_j \cdot \sin(\omega_j t)).$$

From now on we will assume that ℓ was chosen large enough so that the above approximation is—for all practical purposes—an identity. Namely:

$$X'_t = a + b \cdot t + \sum_{j=1}^k (u'_j \cdot \cos(\omega_j t) + v'_j \cdot \sin(\omega_j t)), \text{ for } t = \ell, \dots, (250 - \ell).$$

Next, to accomplish our goal, we would like to use the *Discrete Fourier Transform* (DFT). For this we first need to remove the linear trend in $\{X'_t\}$. To do so, consider the (*causal*) *linear filter*:

$$X''_t := X'_t - X'_{t-1}, \text{ for } t = (\ell + 1), \dots, (250 - \ell).$$

3. Using that $\cos(\alpha - \beta) = \cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta)$ and $\sin(\alpha - \beta) = \sin(\alpha) \cos(\beta) - \cos(\alpha) \sin(\beta)$, show that:

$$X''_t = b + \sum_{j=1}^k (u''_j \cdot \cos(\omega_j t) + v''_j \cdot \sin(\omega_j t)),$$

for suitable constants u''_j and v''_j .^{1 2}

4. If necessary do a literature search about the so-called *Fast Fourier Transform* (FFT). Explain—at a high level—what this algorithm does. Also, report what the following R-command returns: `y = fft(x)`, when `x` is an n -dimensional vector.

¹In practice, the precise relationship between the constants u''_j and v''_j and the original constants u_j and v_j is not relevant. What matters here is that $\{X''_t\}$ is a linear combination of sines and cosines with the same frequencies ω_j 's as the original time series $\{X_t\}$.

²The constant b may be thought of as associated with the frequency $\omega_0 := 0$

5. Explain how to use the `fft` command to determine the following version of the *Discrete Fourier Transform* (DFT) of x :

$$a[k] = \frac{1}{\sqrt{n}} \sum_{t=1}^n x[t] e^{-i(t-1)2\pi(k-1)/n}, \text{ for } k = 1, \dots, n.$$

The rest of this project is devoted to estimate k and the ω_j 's from the dataset in Canvas.

6. Observe that the number of observations from $\{X'_t\}$ becomes smaller as ℓ increases. For instance, if $\ell = 124$ then we will have only two observations from this process! So—in practice—one needs to choose ℓ large enough to effectively damp the noise $\{Z'_t\}$, but small enough to have a sufficient number of observations to estimate parameters robustly. Implement in R a code that assumes a fixed but arbitrary ℓ , and then eyeball the smallest value of ℓ for which the plot of $\{X'_t\}$ looks smooth. Report the value of ℓ you chose, and plot the time series $\{X'_t\}$. Comment on any expected or unexpected behaviour!
7. Implement in R a code that computes the linear filter $\{X''_t\}$, and plot this time series in your report. Comment on any expected or unexpected behaviour!
8. Finally, use the `fft` command to eyeball from the DFT of $\{X''_t\}$ a reasonable value for k , and estimate the frequencies $\omega_1, \dots, \omega_k$. Report these values in a Table. Comment on any expected or unexpected behaviour!