STA457/STA2202 - Assignment 1

Submission instructions:

- Submit *a single PDF file* with your answers to both Theory & Practice parts to A1 on Quercus - the deadline is 11:59PM on Thursday, May 21.
- Your answers to the Theory part can be handwritten (PDF scan/photo is OK).
- Your answers to the Practice part should be in the form of a report combining code, output, and commentary. You can compile your report with RMarkdown (recommended) or another editor (e.g. Word/LaTex).

Theory

1. In this course we work with (weakly) stationary time series. This class of models is closed under linear tranformations, i.e. whenever you take a (non-exploding) linear combination of stationary series, you always end up with a stationary series. For this question you have to prove this result. Consider two *independent* zero-mean stationary series, $\{X_t\}$ and $\{Y_t\}$, with autocovariance functions (ACVFs) $\gamma_X(h)$ and $\gamma_Y(h)$, respectively.

(a) [4 marks] Find the ACVF of the linear combination $Z_t = aX_t + bY_t,\ a, b \in \mathbb{R}$ in terms of the ACVFs of $\{X_t\}, \{Y_t\}$, and show that it is stationary (i.e. only depends on h).

(b) [6 marks] Find the ACVF of the linear filter $V_t = \sum_{j=0}^{p} a_j X_{t-j},\ a_j \in \mathbb{R}$ in terms of the ACVF of $\{X_t\}$, and show that it is stationary.

2. [10 marks] Consider the random walk (RW) series $X_t = X_{t-1} + W_t,\ \forall t \geq 1$, where $X_0 = 0$ and $W_t \sim WN(0, 1)$. Although the series is *not stationary*, assume we treat it as such and calculate the *sample* ACVF $\hat{\gamma}(h)$, based on a sample of size n, as:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h}X_t),\ \ \forall h = 0, 1, \ldots, n-1$$

Show that the *expected value* of the sample auto-covariances are given by

$$\mathbb{E}[\hat{\gamma}(h)] = \frac{(n-h)(n-h+1)}{2n}$$

(*Hint*: the ACVF of X is $\gamma(s, t) = \min(s, t),\ \forall s, t \geq 1$, and the arithmetic series formula is $\sum_{i=1}^{n} i = n(n+1)/2$.)
(*Note*: this illustrates the behavior of the sample ACF of a RW series: it is in fact a quadratic in h, but it behaves very close to linear for the small values of h that appear in the ACF plot.)

Practice

You will work with Statistics Canada's open socio-economic series data. The data are organized by topic in tables, and we will focus on monthly employment numbers by industry (table 14-10-0355-01); see also this brief tutorial. An easy way to access these data directly through R is with the `cansim` library, using "vectors" to identify individual series. **You will be working with employment data for diferent industries and over different time periods, based on the last two digits of your student #, according to the scheme described in the following tables:**

last digit of student #	Industry	Unadjusted	Seasonally adjusted	Trend-cycle
1	Accommodation and food services	v2057828	v2057619	v123355122
2	Agriculture	v2057814	v2057605	v123355108
3	Construction	v2057817	v2057608	v123355111
4	Educational services	v2057825	v2057616	v123355119
5	Forestry, fishing, mining, quarrying, oil and gas	v2057815	v2057606	v123355109
6	Goods-producing sector	v2057813	v2057604	v123355107
7	Information, culture and recreation	v2057827	v2057618	v123355121
8	Manufacturing	v2057818	v2057609	v123355112
9	Public administration	v2057830	v2057621	v123355124
0	Services-producing sector	v2057819	v2057610	v123355113

2nd to last digit of student #	Time period
odd	Jan 1980 to Dec 1999
even	Jan 2000 to Dec 2019

E.g., if your student ID ends in 42, you should use the Agriculture industry data (last digit = 2) over Jan 2000 to Dec 2019 (next-to-last digit = 4 is even). **Beware to use the right data, otherwise you will lose marks**. The following starter code downloads the data for student # ending in 42.

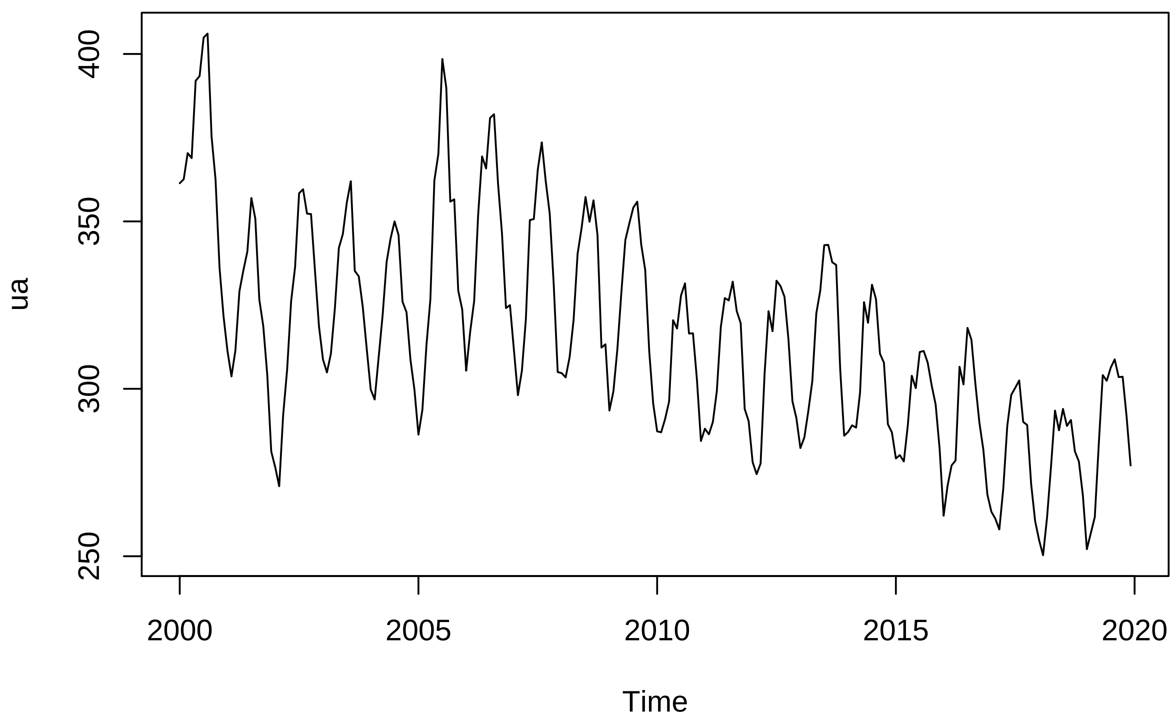
```r
library(cansim)
```

```
## Warning: package 'cansim' was built under R version 3.6.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```r
# unadjusted (raw) series
ua = get_cansim_vector( "v2057814", start_time = "2000-01-01", end_time = "2019-12-01") %>%
  pull(VALUE) %>% ts( start = c(2000,1), frequency = 12)
plot(ua)
```



1. [3 marks] Plot the unadjusted series, its ACF & PACF, and comment on the following characteristics: trend, seasonality, stationarity.

2. [5 marks] Perform a classical *multiplicative* decomposition of the unadjusted series (X_{ua}) into trend (T), seasonal (S), and remainder (R) components (i.e. $X_{ua} = T \times S \times R$):

a. First, apply a *12-point MA* to the raw (unadjusted) series to get an estimate of the trend.
b. Then, use the *detrended* data to estimate seasonality: find the seasonal pattern by caclulating sample means for each month, and then center the pattern at 0 (i.e pattern sum should be 0).
c. Finally, calculate the *remainder* component by removing both trend and seasonality from the raw series. Create a time-series plot of all components like the one below.
(*Hint*: you results should perfectly match those of the `decompose` function, which uses the above process)

3. [2 marks] Statistics Canada (StatCan) does their own seasonal adjustment using a more sophisticated method (namely, X-12-ARIMA). Download the corresponding *seasonally adjusted* series for your industry and time period, and plot them on the same plot with your own seasonally adjusted data ($X_{sa} = X_{ua}/S = T \times R$) from the previous part. The two versions should be close, but not identical. Report the mean absolute error (MAE) between the two versions (StaCan's and yours) of seasonally adjusted data.

4. [5 marks] The library `seasonal` contains R functions for performing seasonal adjustments/decompositions using various methods. Use the following three methods described in FPP for performing seasonal adjustments (you don't need to know their details):

a. X11

b. SEATS

c. STL
   Create seasonaly adjusted versions of your raw series based on each method, and plot them together with StaCan's version. Note that the first two methods (X11 & SEATS) are *multiplicative* by default, and you must use the `forecast` library function `seasadj`, `seasonal`, `trendcycle`, and `remainder` to extract the various components. The last method (STL) however is only *additive*, so you need to take a logarithmic transformation of the data to do the *multiplicative* decomposition, and then transform them back to the original scale for making comparisons.
   Which method gives a seasonal adjustment that is closest to StaCan's, based on MAE?

4. [5 marks] Using StatCan's data (unadjusted, and/or seasonally adjusted, and/or trend-cycle), calculate the *remainder* series (R). Plot R and its sample ACF and PACF, and answer the following questions:

a. Based on these plots, can you identify any remaining seasonality in your series?
b. Comment on the stationarity of the series and propose any further pre-processing.
c. Comment on the (partial) autocorrelations of the series, and propose an appropriate ARMA(p, q) model (i.e. appropriate orders p & q).

5. [10 marks; **STA2202 (grad) students ONLY**] Download employment data *up to April 2020* (the most recent month) for *all* of the above industries, and use them to answer the following question:
   Which industry's employment was hit hardest by the COVID-19 pandemic?.
   You need to back up your answer with valid arguments based on time series techniques, to account for things like seasonality (e.g., you can't simply rank last month's differences in employment numbers). Clearly explain your reasoning and the methods & metrics used for making comparisons.

Acknowldgements:

Thanks to our TA Yang Guo for researching the data used in this assignment.