STAT 4181 - Project 1 - Due April 21, Thursday

Part 1 - R Coding

In this part, you are asked to write a generic R code that will take in a time series named `x` and will:

- Print the ACF and PACF plots of the time series named `x`.
- Fit a series of AR models with $p = 0, 1, 2 \ldots, 10$ and store the AIC of each fit.
- Fit a series of MA models with $q = 1, 2 \ldots, 10$ and store the AIC of each fit.
- Fit a series of ARMA models with $1 \leq p \leq 10, 1 \leq q \leq 10$, and store the AIC of each fit. Note that this requires fitting 100 (10×10) different models and might take a little more time than the other steps. *(Note: For fitting, use the* `arima` *function and set* `order=c(p,0,q)`*.)*
- Find the best among all possible 121 models (11 for AR, 10 for MA and 100 for ARMA) by using AIC, and then finally fit *the best model* and store it as `best.model`.
- Print the ACF and the PACF plots of the residuals of `best.model`.

For this part, only submit your R code. Make sure it is commented properly; the code should include instructions on what it is doing at each stage. As in:

```
#acf and pacf plots
acf(x)
pacf(x)
```

Part 2 - Applications

In this part, you will be using your R code from Part 1 to analyze IMDB ratings of a tv show.

- Download the `tv.R` function from Blackboard.
- Choose 3 tv series with at least 48 episodes and ideally something that has a plot (avoid sitcoms and shows like *The Simpsons* if you can).

- For each tv series:

1. Use the `tv.R` function to obtain the IMDB ratings of that show. You can find an example on how to use the function in `tv-example.R`.
2. Store the ratings in a variable named `x`.

3. Use your code from Part 1 and find *"the best model"* for this tv series. Write a short report that analyzes the output of your R code. This should include the acf and the pacf plots, the parameters of the best model, along with the 2 other models that have the lowest AIC. A couple of sentences on the practical implications of your result (e.g. "the best model is an AR(2) model probably because this show's plot points tend to happen over two episodes") would be appreciated, but are not necessary.

Part 3 - Extensions

In Part 2, you have treated the ratings as a time series with no other known information. In many tv shows, the change in the ratings can also be explained by other factors: shows sometimes get better over time (or worse - think *Lost*); some seasons could be significantly better than the others (e.g. *Community*); there might be an actor/director change that causes a massive shift in the quality (*Community*, again). It is possible that a majority of the variance in the ratings can be explained by these factors.

In this part, you will be using the season numbers as covariates to build a more detailed model.

- Choose one of the three tv shows you have analyzed in Part 2. Try to pick one that has both (i) a clear trend over seasons, (ii) a significant time series component - **the best model for the chosen tv series should not be AR(0) (i.e. white noise)**.
- Fit a least squares model to the time series with the `lm()` function, and use the season numbers as the covariate. Make sure that you `factor` the seasons and remove the intercept term (e.g. `lm(ratings~factor(seasons)-1,data=x)`).
- Report the summary of the fitted model. Are any of the coefficients significant?
- Repeat the analysis in Part 2 on the residuals of the fitted `lm` model. That is, plot the acf/pacf, fit all possible 121 time series models and find the best model.
- Fit a generalized least squares model with the `gls()` function from the `nlme()` package, and for the correlation parameter, use the best model from the previous step:
  `gls(...,correlation=corARMA(COEFS,p=P,q=Q))`,
  where `P` is the chosen AR order, `Q` is the chosen MA order and `COEFS` are your ARMA model coefficients (with length `P+Q`).
- Finally, write a short report (two paragraphs) on the results of the `gls` fit and its implications.