W271 Live Session 10: Vector Auto-Regression

Jeffrey Yau, Devesh Tiwari

3/14/2017

Main topics covered this week

- Regression with multiple trending time series
- Correlation of time series with trends
- Spurious correlation
- Unit-root non stationarity and Dickey-Fuller Test
- Cointegration
- Multivariate Time Series Models: Vector Autoregressive (VAR) model
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference

Readings

CM2009: Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R*. Springer. 2009.

- Ch. 11

SS2016: Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Applications.* EZ Edition with R Examples:

- Ch. 5.3, review Ch. 2.1

HA: Rob J Hyndman and George Athanasopoulos. Forecasting: Principles and Practice:

- Ch. 9.2

Agenda for today

1. SARIMA review (10 mins)
2. Breakout Session 1: (10mins in group, 10 discussion)
3. Group discussion: VAR (15 mins)
4. Breakout Session 2: EDA for VAR (15 mins in group)
5. Breakout Session 3: Building a var (20 mins in groups)
6. Take home

Review of SARIMA

1. Many time-series data in the real world exhibit significant seasonality/seasonal trends. In these data, the value of a given observation is partially a function of it's "season" and it's most proximate prior values.
2. Example: The monthly unemployment rate in the United States exhibits clear seasonal trends; some months always have higher unemployment than others. If we are interetsted in modeling the unemployment rate, then we have to model the dependency of unemployment immediately prior to a given month AND we have to model the dependency of the unemployment rate in the prior year.
3. In order to do this, we can use the SARIMA models, which is often expressed as Arima(p,d,q)x(P,D,Q),m.
4. High level overview of the modeling procedure:
   (a) Conduct EDA to determine d and D.
   (b) Estimate p,q,P, and Q. Note that you can try to estimate these parameters "at once" via 4 nested loops, or you can model the non-seasonal and seasonal components separately. With respect to the latter, remember that you probably will have to "fine tune" your model.
   (c) Evaluate your models based on the following criteria:
      - In sample fit (AIC, BIC)
      - Residual analysis
      - Out of sample fit.
5. Once you have selected a model you like, answer the question at hand (often to produce a forecast).

Breakout Session 1: Compare ARIMA to a SARIMA model

We are going to examine the dataset *cmort* from the astsa library in R. In particular, let's examine the difference between modeling this time-series with an Arima model versus adding a seasonal component. Note, WE ARE NOT GOING TO GO THROUGH THE ACTUAL MODEL BUILDING PROCESS HERE, BUT YOU SHOULD DO THAT AT HOME.

Questions.

1. Examine and interpret the output of each model. What do they say?
2. Examine the residual plots from each model. Comment on any differences you see between the residual plots of each model. Based on these plots, which models do you prefer?
3. Generate and plot a one year forecast for cardiovascular mortality from each model. Comment on the similarities and differenes of each forecast. Note: We did not conduct an out of sample test prior to this step, but you should do that at home. Based on your visual examination alone, which model do you prefer? Why?

```r
rm(list = ls())
library(forecast)
```

```
## Loading required package: zoo
```

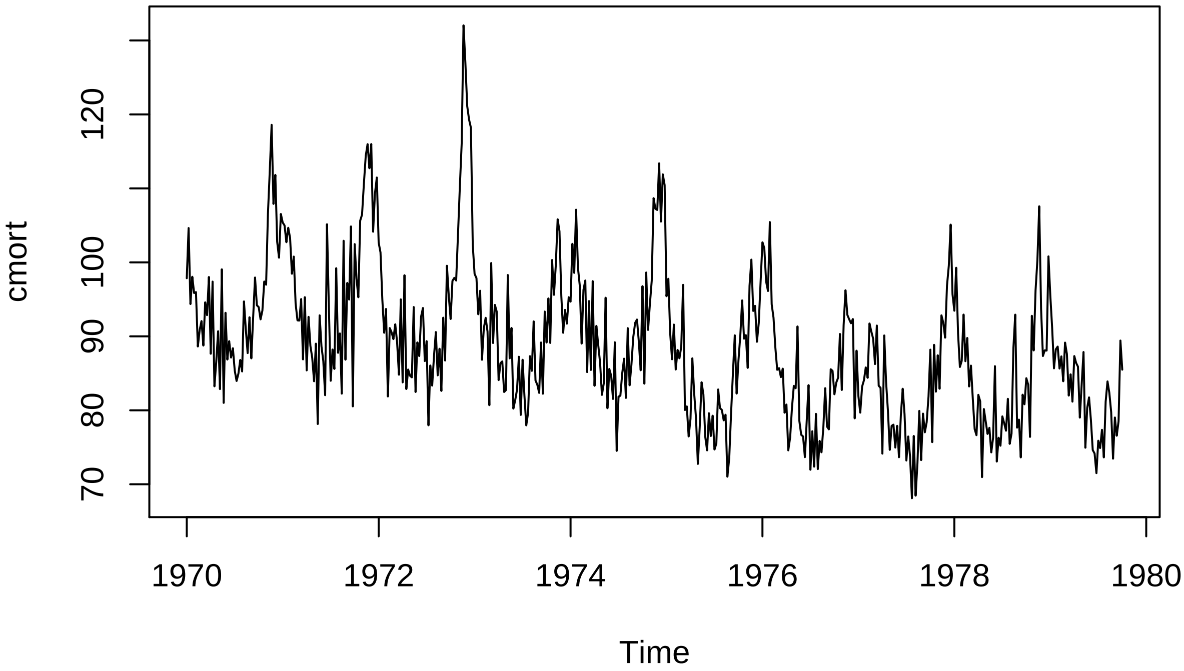
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: timeDate
## This is forecast 7.3
```

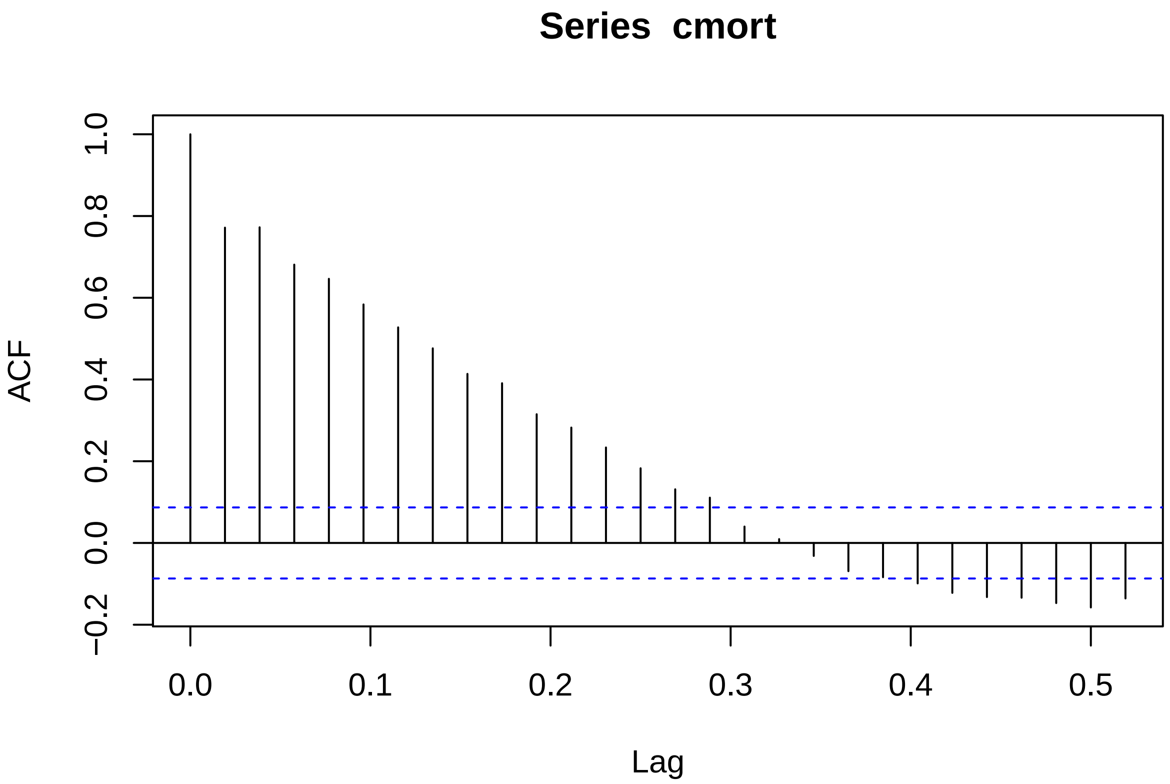
```
library(tseries)
library(astsa)
```

```
##
## Attaching package: 'astsa'
## The following object is masked from 'package:forecast':
##
##     gas
```

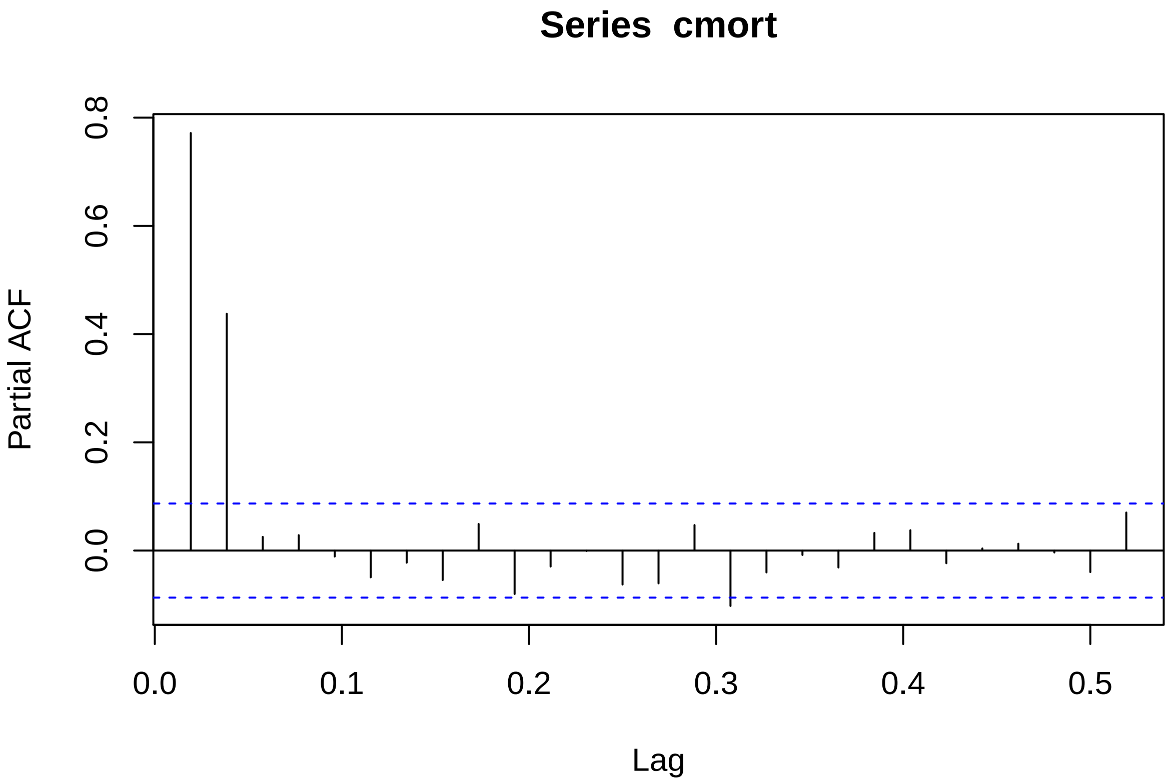
```
# Plot cmort
plot(cmort)
```



```
acf(cmort)
```



```
pacf(cmort)
```



```
# Not stationary and it looks like there is a small, linear trend.
# Let's model this as an Arima(2,0,0) with a trend.
```

```
cmort.ar.drift <- Arima(cmort, order = c(2,0,0),
                        include.drift = TRUE,
                        method = "ML")

cmort.ar.drift
```

```
## Series: cmort
## ARIMA(2,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2  intercept    drift
##       0.4097  0.4226    96.6852  -0.0309
## s.e.  0.0401  0.0402     2.8865   0.0098
##
## sigma^2 estimated as 32.16:  log likelihood=-1600.9
## AIC=3211.81   AICc=3211.93   BIC=3232.96
```

```
# And let's compare this to an Arima()
cmort.seasonal <- Arima(cmort, order = c(2,1,0),
                       seasonal = list(order = c(0,1,1), period = 52),
                       method = "ML")
cmort.seasonal
```

```
## Series: cmort
## ARIMA(2,1,0)(0,1,1)[52]
##
## Coefficients:
##           ar1      ar2    sma1
##       -0.6283  -0.1472  -1.000
## s.e.   0.0463   0.0466   0.111
##
## sigma^2 estimated as 28.7:  log likelihood=-1467.22
## AIC=2942.43   AICc=2942.52   BIC=2958.92
```

Group discussion: Vector Auto-Regression

The astsa library also has time-series data of the weekly temperature in LA during the same period the mortality data were taken. With this time-series data in hand, we can ask some interesting questions:

(1) Is there a relationship between temperature and cardiovascular mortality?

(2) Can we use the temperature data to improve our forecasts of cardiovascular mortality?

Questions

1. Can we use OLS (with DV = *cmort* and IV = *tempr*)? What are the potential shortcomings?
2. What is a unit-root and why do we care about them?

We can use a VAR model if the two variables (and resulting model) are weakily stationary. If they are not stationary, then we have to transform the variables in order to make them stationary. Note, the *VAR* command allows us to incorporate a linear trend and seasonality directly.

Breakout Session 2: EDA portion of VAR

Conduct an EDA on the cmort and tempr data. Be sure to conduct unit-root tests, tests for co-integration (if needed), and examine cross-correlation plots.

```
cmort.train <- cmort[1:458]
cmort.train <- ts(cmort.train, frequency = 52,
                  start = c(1979, 1))
cmort.test <- cmort[459:508]


tempr.train <- tempr[1:458]
tempr.train <- ts(tempr.train, frequency = 52,
                  start = c(1979, 1))
tempr.test <- tempr[459:508]

mortality <- cbind(cmort.train, tempr.train)

### Insert your code here.
```

Breakout Session 3: Building a VAR model

When building a VAR model, you need to figure out how many lags to include (which is the same problem we faced in every other model). R has an automated procedure to do this, in which you select an in-sample criterion to use. Given the importance of forecasting, it is also important to choose models that minimize prediction error. To that end, you should also conduct out of sample tests. Note, that if you have a VAR with n time-series variables, you are actually creating n-models and thus will have to calculate the prediction error across n models. Given that we interested in mortality, just focus on that one for now.

Task 1: Create a VAR model (with a constant and a trend, why?) and no seasonality. Figure out how many lags to include by conducting out of sample tests. Does it make sense to run a 52 step ahead test?

Task 2: Create a VAR model, like you did above, and include a seasonal component.

Take home

Select your favorite VAR model and test it head to head with the SARIMA model from earlier. Which model do you prefer? Why do you think one model is better than the other?