W271 Live Session 10: Vector Auto-Regression

Devesh Tiwari

July 18, 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
2. Group discussion: (10mins in group, 10 discussion)
3. Group discussion: Exploring VAR
4. Breakout Session 1: Conceptual VAR questions
5. Breakout Session 2: EDA for VAR (15 mins in group)
6. Breakout Session 3: VAR model

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).

Group discussion: Vector Auto-Regression

The astsa library has time-series data of the weekly temperature in LA (tempr) and the levels of particulates in the air (part). With these data, we can ask some interesting questions:

(1) Is there a relationship between temperature and air quality?

(2) Can we use the temperature data to improve our forecasts of air quality?

Questions

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

VAR modeling procedure

Similar to the other models we have examined, VAR models require both time-series to be weakly stationary. Unlike the Arima() function in R, the VAR() function does not difference the raw time-series and then re-integrate forecasts. A close examination of the VAR() documentation reveals that the VAR() function can incorporate trends and seasonality in the model. As long as the two time-series are trend stationary and/or stationary after seasonal differences are taken, then the VAR() model can work. Another difference between Arima() and VAR() is that VAR() estimates an equation for each time-series included in the model.

The VAR modelling process is similar to that of ARIMA, with some changes to incorporate the fact that we are now modeling more than one time-series at a time.

Breakout Session 1: Exploratory Questions

(1) In addition to the EDA we perform on univariate time-series, what additional things would you check for when you are conducting an EDA on a VAR model? What questions are you trying to answer about the data before you proceed to the modelling process?

(2) How would you determine the appropriate order of a VAR model? What does this mean?

(3) What are you looking out for when you perform a residuals analysis? How does a residuals analysis here compare to the one you perform in an Arima model?

(4) What is the difference between conducting an out of sample test when building a VAR model versus conducting one for for an ARIMA?

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. Based on your EDA, do you think that there is a positive or negative relationship between temperature and air quality?

```
rm(list = ls())
library(astsa)
library(vars)
```

```
## Loading required package: MASS
```

```
## Loading required package: strucchange
```

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

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

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: urca
```

```
## Loading required package: lmtest
```

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


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

airQ <- cbind(part.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.

(1) Build two VAR models. One with just one lag and another model whose lag length is determined by the VARselect function.

(2) Examine the coefficients of each model. Do they make sense? What is the relationship between temperature and air-quality here?

(3) Examine the residuals of each model. Do you notice anything?

(4) Compare the forecasting accuracy of each model.