Time Series Boot Camp

Day 2: Multivariate Time Series Analysis

Josephine Lukito

March 11, 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

THE J SCHOOL
EST. 1905

Day 2 – What are we doing today?

1. Multivariate Analysis
   - An analysis of multiple time series
   - Vector Autoregression
   - Grainger Causality
   - (Bonus: Impulse Response Functions)

Multivariate Time Series!

Multivariate Time Series

- Many different methods that look at many types of relationships
- We will focus on one frequently used model, but keep in mind that there are many other models that may be useful (a few will be listed at the end)

A note about time and causality

- Time-order is **only one** prerequisite of causality
- Time series analysis allows you to establish **temporal precedence** but does not actually prove **causality**.
- This is important for writing results

- However, experiments often have short-term temporal precedence. Because of the data required for time series, more complex temporal relationships can be identified through these techniques

Vector AutoRegression (VAR)

- Sims, Christopher A. "Macroeconomics and reality." *Econometrica: journal of the Econometric Society* (1980): 1-48.

- Build on the Autoregressive model (the **AR** in ARIMA)
- Highly inductive approach
- This is built on the idea that each variable in a VAR can be modeled as a combination of the variable's past values + past values of other variables + white noise.
- VAR(p) → p is the lag

VAR(p) vs. Structural Equation

VAR

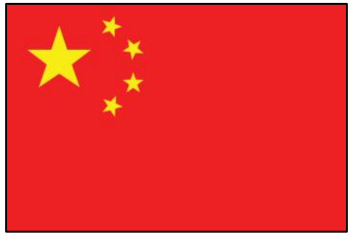
- Theoretically-independent model (in terms of structure)
- Requires time series data
- NOT CAUSAL

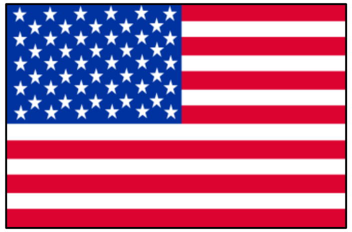
Structural Equation Model

- Reliance on theory to determine functional form
- Does not require time series
- NOT CAUSAL

VAR "Math"

- Let's say, you think Russian, U.S. and Chinese relationships are all related to one another.





- These relationships can be modeled using SEM or VAR!

Let's see this in SEM form!

Let US_t, R_t, and C_t denote the foreign policy behavior of the U.S., Russia, and the P.R.C., respectively. A structural equation model for triangle might be:

$$R_t = \alpha_{10} + \alpha_{11} R_{t-1} + \alpha_{12} C_t + \alpha_{13} US_{t-1} + \epsilon_{1t}$$
$$C_t = \alpha_{20} + \alpha_{22} C_{t-1} + \alpha_{21} R_t + \epsilon_{2t}$$
$$US_t = \alpha_{30} + \alpha_{33} US_{t-1} + \epsilon_{3t}$$

$$\begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix} \sim MVN(0, \Sigma).$$

Let’s see this in VAR form!

$$R_t = \alpha_0 + \overset{\text{Past values}}{\boxed{\sum_{i=1}^{q} \alpha_{1i} R_{t-i}}} + \overset{\text{Past values of other variables}}{\boxed{\sum_{i=1}^{q} \alpha_{2i} C_{t-i} + \sum_{i=1}^{q} \alpha_{3i} US_{t-i}}} + \overset{\text{White Noise}}{\boxed{\mu_{1t}}}$$

$$C_t = \beta_0 + \sum_{i=1}^{q} \beta_{1i} C_{t-i} + \sum_{i=1}^{q} \beta_{2i} R_{t-i} \sum_{i=1}^{q} \beta_{3i} US_{t-i} + \mu_{2t}$$

$$US_t = \gamma_0 + \sum_{i=1}^{q} \gamma_{1i} US_{t-i} + \sum_{i=1}^{q} \gamma_{2i} R_{t-1} + \sum_{i=1}^{q} \gamma_{3i} C_{t-1} + \mu_{3t}$$

Pro’s and Con’s of a VAR Model

Pro’s

- Inductive (a-theoretical)
- Great forecasting capabilities
- Handles exogenous and endogenous variables

Con’s

- Fixed lag for whole model
- Cannot be used for panel data
- Not good for simultaneous relationships
- Requires uniform level of time series aggregation

Preconditions of a VAR

1. All variables are stationary
2. At least 40 data points are needed (100+ is optimal)
3. Uniform data aggregation
4. Your time series must be continuous (regularly spaced)

- You must be able to construct a “tidy” data structure (tidy time!)

Note: Wolfram Alpha

- Search (“Answer”) Engine: https://www.wolframalpha.com
  - [Easter egg: Ask it “What is the answer to life?”]
- Wolfram Alpha will tell you how many days are in between your window of time, inclusive of those dates
  - Try: 6/1/2015 to 11/9/2016
- Wolfram Alpha is also useful for time zone differences, and for converting UNIX time stamps to something readable
  - Another tool: https://www.unixtimestamp.com/

Tidy time series data

- One value, per variable, per time point
  - (observation = time point)
- No duplicate time points
- No missing time points (continuous)

X	date	ru2us	us2ru	usur_conflict	sock_fb	sock_tw	sock_reddit
1	7/2/2015	92	72	86	18	3324	35
2	7/3/2015	78	69	73	0	7098	6
3	7/4/2015	83	60	71	0	4587	2
4	7/5/2015	78	72	61	0	3366	27
5	7/6/2015	98	70	82	3	3985	27
6	7/7/2015	133	156	94	2	4920	2
7	7/8/2015	117	162	121	7	5418	4
8	7/9/2015	125	177	168	7	6241	23
9	7/10/2015	130	140	121	19	6488	20
10	7/11/2015	39	61	59	0	6544	51
11	7/12/2015	54	67	37	0	4338	29
12	7/13/2015	96	118	100	15	8938	24
13	7/14/2015	167	216	120	10	6451	13
14	7/15/2015	165	201	109	9	6382	6
15	7/16/2015	215	241	183	9	4457	20
16	7/17/2015	130	120	125	12	2068	23
17	7/18/2015	65	68	74	0	3645	22
18	7/19/2015	56	69	64	0	3985	13
19	7/20/2015	116	153	138	0	3216	10
20	7/21/2015	107	130	107	7	5699	22
21	7/22/2015	105	113	88	6	7516	26
22	7/23/2015	105	127	71	0	6036	36
23	7/24/2015	115	103	94	1	4210	4
24	7/25/2015	65	53	74	0	4559	14
25	7/26/2015	50	67	80	0	2702	12
26	7/27/2015	96	101	102	3	5795	16

Steps to a VAR Analysis

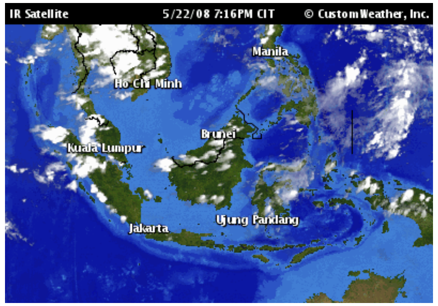
Model specification | Construction | *Analysis and Model Checking*

1. **First-difference any integrated univariate time series**
2. **Determine order, and endogenous/exogenous variable**
3. **Identify the optimal lag**
4. Construct a VAR model
5. *Run Granger Causality Tests*
6. *Run Impulse Response Functions*
7. *Check residuals of model*

Exogenous vs. Endogenous Variables

- **Endogenous** – A variable that can be explained by the relationship of other variables in a model

- **Exogenous** – A variable that is cannot be explained by the relationship of other variables in a model
  - It may be explained by things **outside** of the model, but it's not explained by things **in the model**.



Determining an appropriate lag

- Use information criteria (AIC or BIC) to determine the appropriate lag
- Select the lag with the largest number

```
library(tsDyn)

tot <- lags.select(var_endo, lag.max = 8) #I like this one because I rely on the BIC most often
tot
```

Dataset of VAR variables

Tests for 1-8 lags

```
## Best AIC:  lag= 4
## Best BIC:  lag= 2
## Best HQ :  lag= 2
```

```
tot$BICs
```

(Shows BIC for models with lags of 1 to 8)

```
##         lag=1    lag=2    lag=3    lag=4    lag=5    lag=6    lag=7 lag=8
## r=0 3065.461 3057.89 3084.049 3097.651 3120.508 3150.013 3178.139    NA
```

A lower number is better

VAR Results

In this [fractionally integrated] VAR, Habel focuses on *The New York Times* as the dependent variable. He notes, “the *Times* responds to the Democratic party median in the House and the Senate at a statistically significant level” (p. 270)

Habel, P. D. (2012). Following the opinion leaders? The dynamics of influence among media opinion, the public, and politicians. *Political Communication*, *29*(3), 257-277.

Table 2
Fractionally integrated vector autoregression of the *New York Times* and Democrats in Congress, 1952–2006

Dependent variable	Explanatory variable	1	2
House Democrats$_t$	House Democrats$_{t-1}$	−.54*	−.54*
		(.14)	(.14)
	Senate Democrats$_{t-1}$	.14	.14
		(.17)	(.17)
	New York Times$_{t-1}$	−.04	−.04
		(.07)	(.07)
	Unemployment rate$_t$		.96
			(.71)
	Inflation rate$_t$		−.04
			(.38)
	Constant	1.57	1.57
		(2.29)	(2.24)
Senate Democrats$_t$	House Democrats$_{t-1}$	−.15	−.15
		(.12)	(.12)
	Senate Democrats$_{t-1}$	−.22	−.22
		(.15)	(.15)
	New York Times$_{t-1}$	−.05	.04
		(.06)	(.06)
	Unemployment rate$_t$		.31
			(.62)
	Inflation rate$_t$		.03
			(.33)
	Constant	7.92*	7.95*
		(1.95)	(1.97)
New York Times$_t$	House Democrats$_{t-1}$	.72*	.72*
		(.29)	(.29)
	Senate Democrats$_{t-1}$	.56†	.56†
		(.34)	(.34)
	New York Times$_{t-1}$	.38*	.38*
		(.14)	(.15)
	Unemployment rate$_t$		−.22
			(1.47)
	Inflation rate$_t$		.09
			(.79)
	Constant	9.88*	9.95*
		(4.62)	(4.67)

Note. $N = 42$. Standard errors are in parentheses. Durbin Watson Model 1 = 2.15, Model 2 = 2.20.
*$p < .05$; †$p < .10$.

Why do people gloss over the VAR model?

- VAR model results are in a reduced form, so the coefficients are difficult to interpret (the coefficient is only one part of the model—there is also a function for each endogenous variable)
- VAR models are **long** (the more variables you include, the longer your chart gets)

- For more, please read: Qin, D. (2011). Rise of VAR modelling approach. *Journal of Economic Surveys*, 25(1), 156-174.

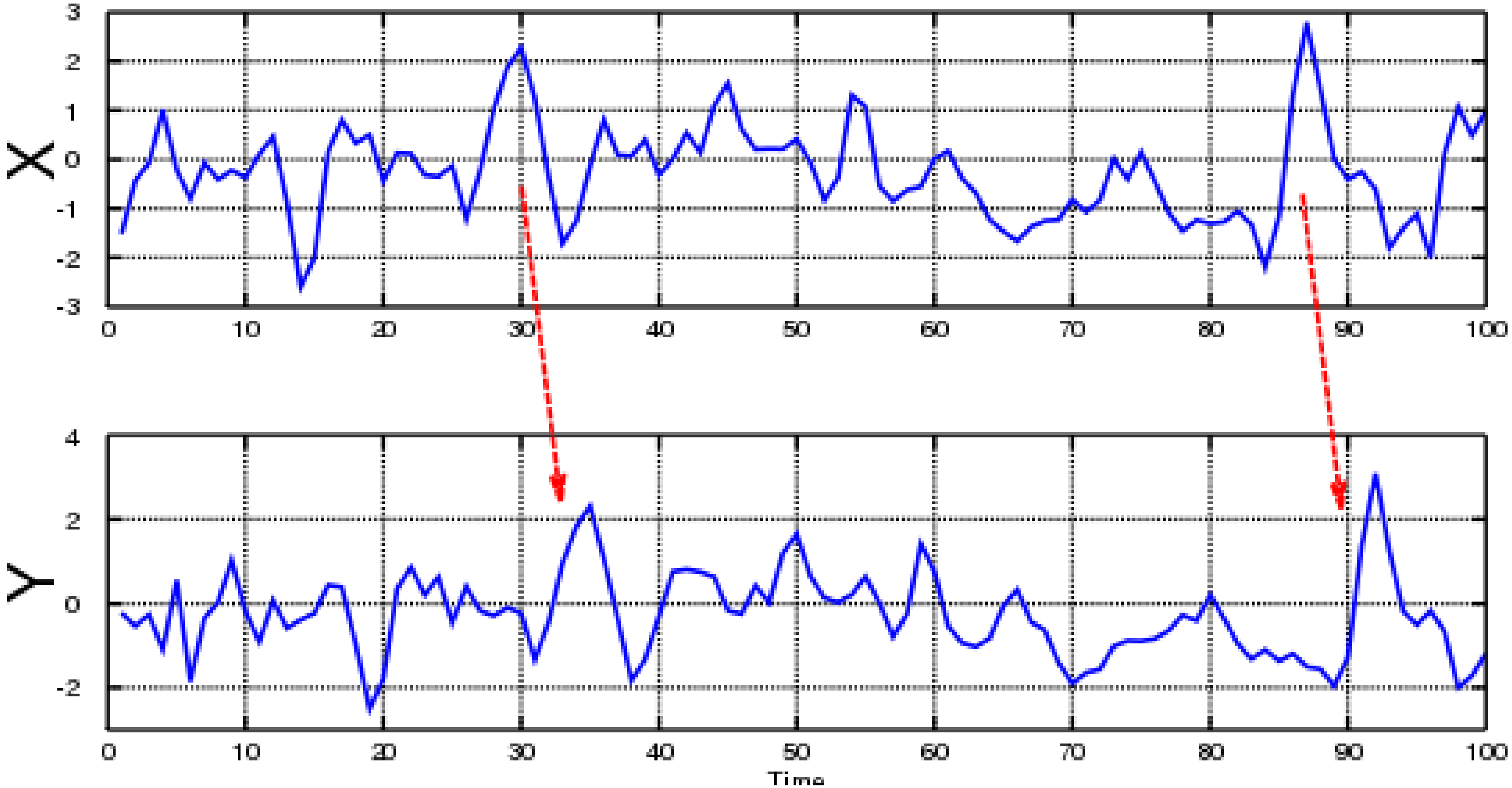
What do people use instead?

1. Granger Causality
2. Impulse Response Functions

Granger Causality

- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.
- Relies on the VAR model
- When *X* "granger causes" *Y*, it means *Y* can be better predicted (or forecasted) with the inclusion of *X* than by *Y* itself (Diebold, 2001)

- **Granger causality is NOT <u>causality</u>**
- Because this is not an experiment, we cannot control "everything." Therefore, omitted variables could still affect the causality hypothesis



Source: https://blog.projectpiglet.com/2018/01/limits-of-granger-causality/

Limits of Granger Causality Tests

- Granger causality tests are really focused on time-lagged relationships. In other words, things that occur simultaneously cannot be “teased out” to see if one Granger causes another.

- Both *X* and *Y* have to be endogenous variables

Granger Causality Results

	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4
Public→Public					Media→Media				Policy→Policy			
defense	**7.11*** **	**-1.99***	**2.13***	-1.75+	**2.33***	**2.15***	1.03	-1.36	**2.29***	**3.84*** **	-1.20	-0.56
government	**2.76** **	1.81+	0.75	**-2.27***	**2.07***	0.52	0.82	1.79+	1.16	1.28	-0.37	1.95+
international	**4.55*** **	-1.31	1.49	0.83	**4.62*** **	0.22	0.01	-0.87	**3.07** **	0.19	0.14	**3.11** **
Crime & law	**5.75*** **	-0.98	1.24	-0.48	**2.64** **	0.83	-0.17	0.30	1.78+	1.62	0.68	1.44
Policy→Public					Public→Policy							
defense	-1.01	0.69	**1.97***	**-1.99***	-0.68	-0.11	0.22	0.50				
government	0.71	-1.69+	-0.49	-0.89	0.55	-0.67	1.16	-0.73				
international	-1.78+	-0.79	1.73+	1.13	-0.89	0.05	0.52	-1.82+				
Crime & law	-1.75+	0.75	0.01	1.26	0.39	0.53	-0.16	-1.04				
Media→Public					Public→Media							
defense	1.15	1.22	-1.19	-1.08	1.14	0.12	**2.15***	**-2.12***				
government	0.47	-1.10	-0.39	**2.17***	0.79	-0.56	-0.62	-1.22				
international	**2.02***	**-2.34***	0.73	0.87	0.16	-0.46	-0.20	1.50				
Crime & law	1.08	-0.29	0.42	-1.65+	1.58	-1.89+	1.86+	0.44				
Media→Policy					Policy→Media							
defense	-0.38	0.49	-1.34	1.73+	1.59	0.25	**-2.06***	0.26				
government	1.01	-0.68	0.85	-0.22	0.62	-1.25	0.98	-1.11				
international	-0.38	**2.42***	-1.83+	0.46	0.87	-0.07	**-3.17***	1.89+				
Crime & law	0.56	0.07	0.26	0.87	0.56	-0.30	0.16	-0.52				

Note: ***$p<.001$, **$p<.01$, *$p<.05$, +$p<.10$.
Symbol "→" means "Granger-causes." The analysis is trivariate Vector Autoregression containing four lags (L1, L2, L3, L4) of the dependent variable and four lags of each of the two predictor variables. Block Z tests determine whether the block of four lags of the predictor variable of interest improves the model fit. There were 46 observations in the series, running from 1951 to 2003.

Tan, Y., & Weaver, D. H. (2007). Agenda-setting effects among the media, the public, and congress, 1946–2004. *Journalism & Mass Communication Quarterly*, *84*(4), 729-744.

Granger Causality and Mass Comm

- **<u>Agenda setting</u>**
  - Tan, Y., & Weaver, D. H. (2007). Agenda-setting effects among the media, the public, and congress, 1946–2004. *Journalism & Mass Communication Quarterly*, *84*(4), 729-744.
  - Meraz, S. (2011). Using time series analysis to measure intermedia agenda-setting influence in traditional media and political blog networks. *Journalism & Mass Communication Quarterly*, *88*(1), 176-194.
- **<u>Content + secondary data</u>**
  - Groshek, J. (2011). Media, instability, and democracy: Examining the Granger-causal relationships of 122 Countries from 1946 to 2003. *Journal of Communication*, *61*(6), 1161-1182.

More Readings on Granger Causality

- http://www.scholarpedia.org/article/Granger_causality
- https://www.r-bloggers.com/chicken-or-the-egg-granger-causality-for-the-masses/
- Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, *105*(1), 131-159.

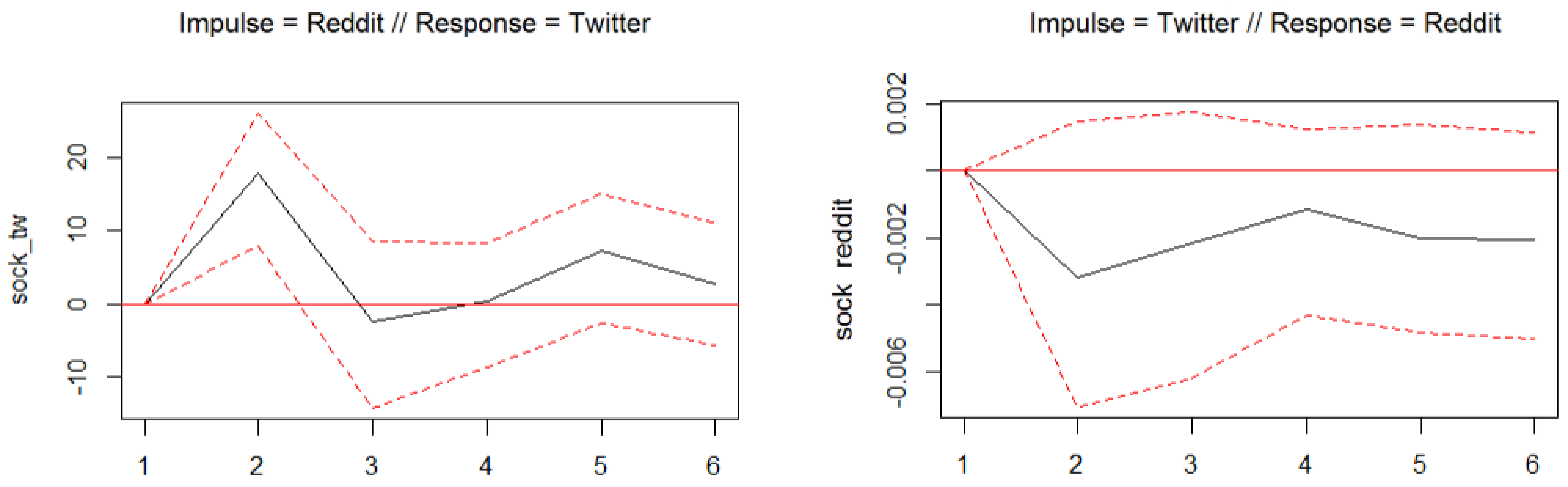
Impulse Response Functions

- Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics letters*, *58*(1), 17-29.
- Used to identify what happens to one variable (Y) when another variable in the model (X) is “shocked” over *n* lags.
- Both X and Y have to be endogenous variables

- In mass comm research, these are less common compared to VARs and Granger Causality tests

IRF Example

Figure 3: Impulse Response Functions between IRA activity on Reddit and Twitter



More Readings on IRFs

- The order of your VAR influences your interpretation of the Impulse Response Function

- https://www.r-econometrics.com/timeseries/irf/
- https://www.ssc.wisc.edu/~bhansen/460/460Lecture25%202017.pdf
- http://economia.unipv.it/pagp/pagine_personali/erossi/dottorato_svar.pdf

Other Multivariate Models

- Error Correction Models
  - A different way of dealing with cointegrated time series
  - Soroka, S. N., Stecula, D. A., & Wlezien, C. (2015). It's (change in) the (future) economy, stupid: economic indicators, the media, and public opinion. *American Journal of Political Science*, *59*(2), 457-474.
- ARCH/GARCH
  - Accounts for change in variance over time
- Dynamic Conditional Correlations
  - Lebo, M. J., & Box-Steffensmeier, J. M. (2008). Dynamic conditional correlations in political science. *American Journal of Political Science*, *52*(3), 688-704.

Plan for tomorrow (Coding Day!)

- In addition to re-applying some of our new ARIMA skills to new data, we will be focusing on the following topics:

1. Jordan & Jo’s favorite packages for time series analysis
2. A bit on data wrangling for VAR
3. Constructing a VAR model (+ Granger causality tests and IRFs)
4. Post-hoc Univariate and Multivariate Model Checking
5. Data visualization (static and interactive)