Time Series Analysis

SVAR to Bayes to BVAR

Michael Colaresi
colaresi@msu.edu

Outline

SVAR

Example

Bayes vs. Frequentists

Subjective Probability

Types of Priors
- Hyper-parameters
- Sims-Zha and Non-stationarity

Review Guide

SVAR

Time to bring structure – if you have it – to VAR.

$$A_0 y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t$$

$y_t = (y_{1t}, y_{2t}, \ldots, y_{mt}); [m \times 1]$
$y_{t-i} = (y_{1t-i}, y_{2t-i}, \ldots, y_{mt-i}); [m \times 1]$
$u_t = (u_{1t}, u_{2t}, \ldots, u_{mt}); [m \times 1]$
A_i are $m \times m$ coefficient matrices

- Notice the dimensions
  - $(m \times m)(m \times 1) = (m \times m)(m \times 1) + \ldots + (m \times 1)$
  - Output will be $m \times 1$ on both sides.

A Knot

- Structural relationships are in A_0

  $A_0 y_t$

  $A_0; [m \times m]$ has elements, $\begin{bmatrix} A_{11} & A_{12} & \cdots \\ A_{21} & A_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$

  If m=2, $A_0 y_t = \begin{bmatrix} A_{11} y_{1t} + A_{12} y_{2t} \\ A_{21} y_{1t} + A_{22} y_{2t} \end{bmatrix}$
- A_{ii} are error variances
- $A_{ij}, \forall\ i \neq j$; tells us that a change in series j has a simultaneous effect on series i.
- Unidentified for all 4 unknowns.
- We can put restrictions on A_0 to identify the system

Example

Two variable model, recursive structure, for the Israeli-Palestinian model.
columns are shocks; rows are responses

	I2P	P2I
I2P	A_{11}	A_{12}
P2I	0	A_{22}

OR,

	I2P	P2I
I2P	A_{11}	0
P2I	A_{21}	A_{22}

Think about these as regressions if there were no dynamics:

$$
\begin{aligned}
y_{I2P,t} &\sim A_{11} \\
y_{P2I,t} &\sim A_{22} + A_{21} y_{I2P}
\end{aligned}
$$

Example: Adding US

Now, a 3 Variable model ($m = 3$),

	I2P	P2I	US
I2P	A_{11}	0	?
P2I	A_{21}	A_{22}	?
US	A_{31}	A_{32}	A_{33}

Make recursive with the US first.

	US	I2P	P2I
US	A_{00}	0	0
I2P	A_{10}	A_{11}	0
P2I	A_{20}	A_{21}	A_{22}

Which has the simultaneous regression form:

$$
\begin{aligned}
Y_{US} &\sim A_{00} \\
Y_{I2P,t} &\sim A_{11} + A_{10} Y_{US} \\
Y_{P2I,t} &\sim A_{22} + A_{21} Y_{I2P} + A_{20} Y_U S
\end{aligned}
$$

Restrictions

Note that to identify the 3 variable model, we need 3 simultaneous restrictions (these are zeros).

What we would like is to place a non-zero ABOVE the diagonal, means a non-recursive model (estimate A_{12} an A_{21} simultaneously).

	US	I2P	P2I
US	A_{00}	0	0
I2P	A_{10}	A_{11}	A_{12}
P2I	A_{20}	A_{21}	A_{22}

But that is unidentified (not enough zeros).

Non-recursive model

What we can do is zero out a different cell.

	US	I2P	P2I
US	A_{00}	0	0
I2P	A_{10}	A_{11}	A_{12}
P2I	**0**	A_{21}	A_{22}

Assumption, Palestinian authority does not act towards Israel IMMEDIATELY in reaction to what US does to Israel (known as a triangular restriction). We have purchased a coefficient for a zero restriction (no free lunch).

Regression structure:

$$
\begin{aligned}
Y_{US} &= A_{00} \\
Y_{I2P,t} &= A_{11} + A_{12} Y_{P2I} + A_{10} Y_{US} \\
Y_{P2I,t} &= A_{22} + A_{21} Y_{I2P}
\end{aligned}
$$

Dynamics

Now add dynamics back as well as shocks,

$$\begin{aligned} A_{00} Y_{US,t} &= D_1 + u_{1t} \\ A_{11} Y_{I2P,t} + A_{12} Y_{P2I} + A_{10} Y_{US} &= D_2 + u_{2t} \\ A_{22} Y_{P2I,t} + A_{21} Y_{I2P} &= D_3 + u_{3t} \end{aligned}$$

where $D_k = \Sigma_1^n \Sigma_1^p A_{kn,p} Y_{n,t-p}$
stack everything back in matrix form, and assume SVAR(1),
$A_0 y_t = A_1 Y_{t-1} + u_t$
multiply each side by A_0^{-1}
$y_t = A_1^* Y_{t-1} + u_t^*$ where, $A_1^* = A_1 A_0^{-1}$ and $u_t^* = u_t A_0^{-1}$
This tells us that we can recover the structure from the $m \times m$ error-covariance matrix

To estimate, first estimate reduced form VAR. Then solve for the unknowns in A_0 using the residual covariance matrix.

Rewind

left hand-side, $A_0 y_t$ is $m \times m$ $\begin{bmatrix} A_{00} Y_{US} & 0 & 0 \\ A_{10} Y_{US} & A_{11} Y_{I2P} & A_{12} Y_{P2I} \\ 0 & A_{21} Y_{I2P} & A_{22} Y_{P2I} \end{bmatrix}$
can factor into two matrices: Y_t; $[m \times 1]$, which is known, $\begin{bmatrix} y_{US} \\ y_{I2P} \\ y_{P2I} \end{bmatrix}$

and A_0; $[m \times m]$, which is mix of "knowns"(constraints) and unknowns.
$\begin{bmatrix} A_{11} & 0 & 0 \\ A_{10} & A_{22} & A_{12} \\ 0 & A_{21} & A_{33} \end{bmatrix}$
The estimates of the $A_{ij}, \forall\ i, j \leq m$ defines the structure. Lets look at how to do this and some estimates.

Notes

- Can not just put any restrictions on A_0. Need to have unique information in each row. For example this A_0 is unidentified,

	US	I2P	P2I
US	A_{00}	0	A_{02}
I2P	0	A_{11}	A_{12}
P2I	0	A_{21}	A_{22}

- You can have an over-identified model (add "extra" restrictions).
- There are other types of restrictions. Can restrict error correlations, replace Iu_t, where I is the identity matrix, with $B_0 y_t$, which you would restrict in the same way as A_0. Restriction here are on correlations among unobserved shocks, not on observed responses to shocks (similar though).
- Economists also tend to sometimes use long-term restrictions to identify a model. These are restrictions on the equilibrium behavior or series.
- Instead of zero restrictions you can also have sign restrictions. This has been implemented in some Matlab toolkits in a Bayesian framework.

Bayes vs. Frequentists

Bayesian Talking Points

- Use Probability to Quantify Uncertainty ($p(MODEL|DATA)$)
- Subjective probability (Its all in your head)
- Prior information
- Explicit rule for updating beliefs as new information become available.
- Mantra: The posterior is proportional to the prior times the likelihood (Translation: Data is only one part of statistical inference.)

Frequentist Answer

- Do no such thing, for shame!
- Uncertainty is in data, not in your head.
- Mantra- Inference is proportional to data and that is the data, do not mess with the data. Stop it....don't touch that data!

Reinterpreting Inference

- We are shown, $y \sim p(y, \theta)$
- We want to know θ, or more specifically, we want $p(\theta|y) \leftarrow$ The distribution of the unknown parameter conditional on the observed data.
- Frequentists and Bayesians disagree about how to make inference.

Franklin's History of Statistics

- In the beginning, there was Bayes (1763)
- But Bayes we too difficult, so Legendre and Gauss (1804-809) said, minimize squared error, and that was easy.
- But it wasn't general and was not rooted in probability theory.
- So Fisher (1922-25) said maximize the likelihood, and it was good again.
- But Savage and de Finetti (1950-1970) said probability is in your head.
- And Gelfand and Smith (1990) pointed out ways to make what was hard (Bayesian probability models) much less difficult.
- And it was good.

In all Likelihood

Fisher is the poster-child for frequentists.

- A "likelihood" is proportional to the density function of the data, given the parameters, with the data treated as fixed and the parameters varying.

$$L(\theta|y) \propto p(y|\theta) \tag{1}$$

- BUT THIS IS NOT, $p(\theta|y) =$ inverse probability.
- The value of θ which maximizes the likelihood of a particular sample y is the mle.
- ML estimators have nice asymptotic properties: Take MLE class if you have not already.

Fisher hated Bayesian ideas:

- Argued that statistical inference should be based on the data, and the data only.
- Main critique: Bayesian use of prior information causes the elegance of likelihood to be lost (things become messy).
- Information should not be produced out of thin air.
- Quote: "[Bayesianism] like an impenetrable jungle arrests progress towards precision of statistical concepts."

Fisher's Folly

But it turns out Fisher still wanted to work with a full probability model (look at $p(\theta|y)$).

- Tried to create "fiducial inference", which was an attempt to do Bayesian statistics without prior information.
- Turns out his solution either was wrong (Efron 1998) or was equivalent to simple Bayesian methods (Lindley 1985).

MCMC

Where Fisher failed (and Bayes did not have a computer), Physicists gave the gift of MCMC.

- Markov Chain Monte Carlo techniques (another class).
- Allows us to sample for very complex probability distributions.
- Makes Bayesian inference practical.
- No longer must have likelihood's elegance to get answer.

What is all the fuss about

- A joint distribution is the product of a conditional and a marginal probability:
  $p(\theta, y) = p(\theta)p(y|\theta)$
  and also we can write,
  $p(\theta, y) = p(y)p(\theta|y)$
  What we need to make probabilistic inference is the last term ($p(\theta|y)$). So,

$$
\begin{aligned}
p(\theta|y) &= \tfrac{p(\theta,y)}{p(y)} \\
&= \tfrac{p(\theta)p(y|\theta)}{p(y)}
\end{aligned}
$$

What this means

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2}$$

We can write this as:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{3}$$

The Mantra (again): The Posterior is proportional to the prior times the likelihood.

Prior?

Priors:

- How can we write one down ($p(\theta)$=?)
- θ is fixed in the population, how can we have a probability distribution for something that is fixed?
- Even if you deal with previous questions, how do we know?
- Frequentists have just left the building.
- Bayesian have just started buying fast computers.

Subjective Probability

- Can probability be subjective (Savage and de Finetti)?
- What does this mean?
- Objective Probability- Neyman and Pearson
  - Coin flip
  - What is probability of heads?
  - Hypothetical repetition of DGP (flip) large/infinite number of times.
  - Data (runs of heads and tails) come from underlying truth (θ), and underlying truth comes from ? (answer depends on your religion).
  - We learn about underlying truth through data (y)-hopefully lots of it.
  - Therefore we have $p(y|\theta)$, but NOT $p(\theta)$, there is no probability statement about a constant (there is just θ!!), and hence we do not have $p(\theta|y)$.

Subjective Probability Continued

- Bayesians see probability very differently.
- Probability is subjective.
- My individual likelihood that a coin is fair.
- My individual beliefs about the coin flip would be defined in ($p(\theta)$).
- For frequentists, θ is fixed. For Bayesians our subjective knowledge of θ is NOT fixed.

Confidence intervals vs. Credible Intervals

- Frequentists say: "If we repeated this experiment a large number of times, the true influence of this lecture on sleep duration would be between .4 and .6 hours 95% of the time."
- Bayesians would say: " There is a 95% chance that the this lecture leads to sleep durations of between .4 and .6 hours."
- Which do you prefer?

From Prior to Posterior

Prior interacts with the data to give you $p(\theta|y)$.

- A vague (highly uncertain prior) will let the data dominate the prior.
- The more precise the prior, the more the posterior will be anchored to the prior (more difficult for data to change our mind).
- Given precise enough data, the likelihood dominates any reasonable prior.

Bayesian ideas in Common Frequentist Statistics

- Specification: What variables go in the model, which ones do not (0/1 beliefs?)
- Functional form: What is the relationship between variables if non-zero (are we sure things are linear)?
- One-tailed tests: Obviously, there is some prior expectation.
- Missing values: Drop them?
- Confidence intervals: What we want them to mean, but they do not mean (probability statements about $\theta|y$).

Advantages of Bayesian Models

- Overt and clear assumptions (like formal theory for data)
- Systematic incorporation of previous knowledge (DPG outside of data)
- Probability interpretation of all quantities of interest!!!
- Ability to update as new data come in.
- Missing data is naturally handled (priors over what would be there).
- Complex models become tractable.
- Latent variables easily and intuitively incorporated.
- Post-estimation we have a posterior distribution of our parameters given the data. Easy to forecast, predict and extend.

Disadvantages of Bayesian Models

- Requires non-trivial knowledge of probabilities and distributions.
- No more "point-and-click", more like, "think-and-program".
- Must convince people about priors to get them to take your posterior inference seriously.
- Estimation is still hard for many models and time-consuming.

Bayesians for Time Series

Notice that advantages of Bayes are greater for time serial data than many other applications.

- Sequential updating
- prior knowledge about out of sample behavior (long memory, unit root, stationary).
- Do not need to choose one model, but can have a probability distribution over models.
- CAN COMPARE MODELS OUT OF SAMPLE (Forecasts), helps to sell priors (useful priors should produce useful forecasts).

Simple Example

Imagine an election where a government will either be re-elected or kicked out of office. We want to look at support for government, as well as who will win.

- Need Prior for proportion of people that support government. This will be our $p(\theta)$.
- As campaign begins we get polling data.

Example II

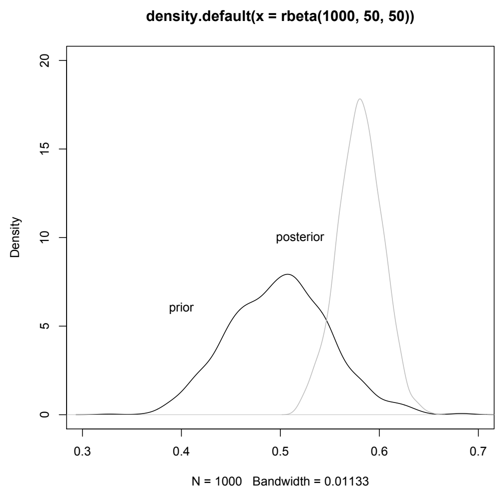
- Assume that each poll consists of N voters, and y of whom say they support the government ($N - y$ are for the opposition, thus no don't knows, Nader-ers or Perot-ites).
- Assume (unrealisticly) that voters are indistinguishable in the probability of voting for the government. Then we have y resulting for a series of binomial trial.
- Prior needs to be bounded between 0 and 1. This can be the Beta distribution – very flexible and bounded.
  - Suppose my belief is that the Government only has the support of half of the population ,and I think that estimates has a standard deviation of .07. This is a B(50,50) distribution.

Data come in

We now observe a poll with 400 respondents, 240 of whom (60%) say they will vote for the government.

- A Beta distribution times a Binomial distribution (Data) gives you a Beta distribution (this is know on conjugacy). Thus, the posterior will be Beta.
- Specifically, we will have $B^{posterior}(\alpha, \beta)$, where $alpha = 50 + y = 290$ and $\beta = 50 + N - y = 210$, thus $B^{posterior}(290, 210)$
- Mean of B(290,210) is $\frac{\alpha}{\alpha+\beta}$ $210/400 = .58$.
- Variance is slightly more complicated but is .0221.
- can also just simulate in R (rbeta(1000,290,210)).

Posterior

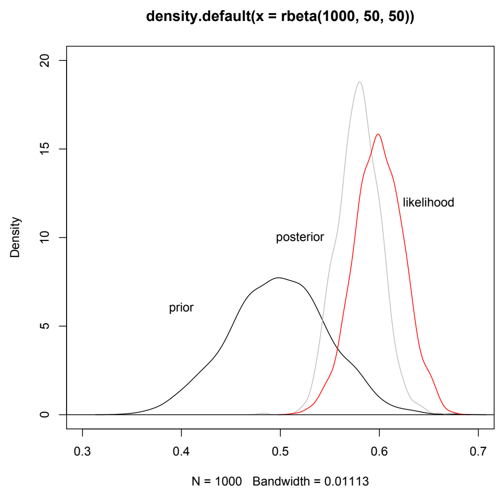


Frequentist Analysis

No prior. Just the data.

- $\hat{\theta} = .60$ and $se_{\hat{\theta}} = \sqrt{.6 \times .4/400} = .0245$
- Compare to Bayesian estimate this is bigger and slightly more uncertain. Bayesian is smaller and slightly more certain.
- Prior draws the data towards itself.

Bayesian Inference



Winning

But we want to know who is going to win

- This is NOT NOT NOT .58 or .60. Those are expected votes.
- Frequentist answer? - Given a value, overlap .5, blah, blah.
- Bayesian—easy. Draw a large number of samples from the posterior distribution (in this case (B(290,210))) and just count the proportion of times that the outcome is greater than .5.

```
 posterior<-rbeta(10000,290,210)
plot(density(posterior))
winner<-ifelse(posterior>.50,1,0)
table(winner)/length(winner)
```

Bayesian Example: Take II

You have just accidentally dropped a nuclear weapon off the coast of North Carolina. It was unarmed, but you would like to get it back (before someone else claims it). What is the frequentist approach to solving this problem?

- Scanner, radar, patrol boats, coast guard, etc.
- Search whole area until it is found.
- Possibly responsible for the destruction of a major US city if someone else discovers bomb first.

What about the Bayesian way of solving the problem?

- Use last coordinates of airplane to start search.
- Use scanners, radar, patrol boats, coast guard, etc.
- Possibly responsible for saving a major US city.

This is an informative prior. You want to add information in this case.

Priors

- Other types of Priors:
  - Uninformative Prior - Think uniform dist. You do not want to add any information to the data.
  - Conjugate Prior- A prior distribution that when multiplied by the distribution of the data will yield a posterior distribution that is the same distribution as that prior. (Example, Beta times a Binomial gives you a Beta; Dirichlet times a Multinomial gives you a Dirichlet; Gamma times a Poisson gives you a Gamma).
  - Reference Prior - prior settled on by many people, perhaps a whole discipline (sometimes noninformative, but not always).

BVARs- The First Generation

BVAR from Minnesota (Doan, Litterman and Sims)

- Prior on the relationship between lags in a VAR
- Smoothes the IRFs out.
- Priors on coefficients in the reduced form

Justification:

- A prior that most time series are best predicted by their mean or their values in the previous periods.
- For non-stationary data this means that the data are first-order integrated perhaps with drift (deterministic constants) or that the first differences of each series are unpredictable.

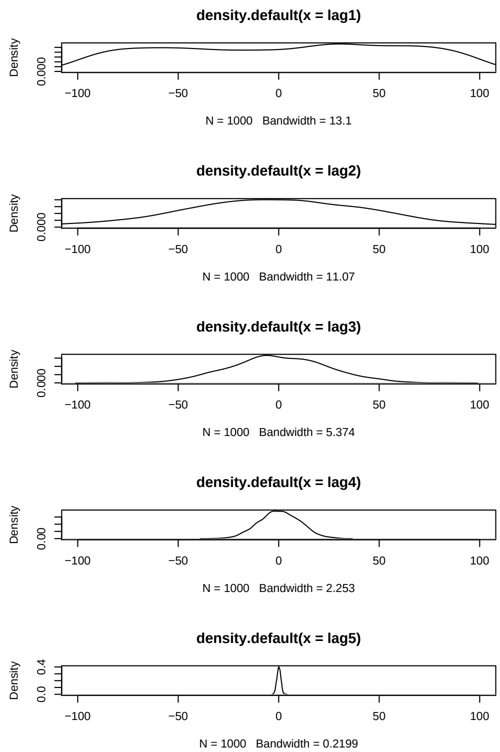
What happens

More specifically,

- Prior is that all coefficients except the coefficient on the first own lag of the dependent variable have mean zero and that certainty about this belief is greater the more distant the lag of the variable to which the coefficient applies

```
lag1<-runif(1000,-100,100); lag2<-rnorm(1000,0,50);
lag3<-rnorm(1000,0,25); lag4<-rnorm(1000,0,10);
lag5<-rnorm(1000,0,1); par(mfrow=c(5,1));
plot(density(lag1), xlim=c(-100,100))
plot(density(lag2), xlim=c(-100,100))
plot(density(lag3), xlim=c(-100,100))
plot(density(lag4), xlim=c(-100,100))
plot(density(lag5), xlim=c(-100,100))
```

Minnesota Prior



Whose Prior?

Problems:

- One of the key features of this prior is that it treats the variance-covariance matrix of the reduced form residuals as diagonal and fixed.
- In addition, it does not embody any beliefs an analyst might have about how the prior distribution of the variance-covariance matrix of residuals is related to the prior distribution of the reduced form coefficients (What if there was a huge simultaneous correlation between two variables? Would our prior on the first lag remain the same?).
- Posterior distributions were unworkable (really complicated)

Enter the Princeton Prior: BVAR The Sequel

A relatively new prior for BVAR models has been offered by Sims and Zha. This prior:

- Sets up priors on structural model.
- Includes beliefs about non-stationarity and cointegration
- Idea: If innovations in two equations are highly positively correlated, then we would expect the lagged behavior to be positive and meaningful also. Connection between innovations and reduced form (through the structural model).

Properties

Nice properties:

- Can extend to SVAR (unfortunately titled B-SVARs)
- Can represent non-stationarity/near-non-stationarity
- Can include more variables than in VAR (or SVAR), thus can represent more complex systems.
- Natural extension to impulse response function error bands.
- Forces you to think.

Not so nice-properties:

- Can take a while to estimate (minutes to days).
- Forces you to think.

Hyper-parameters

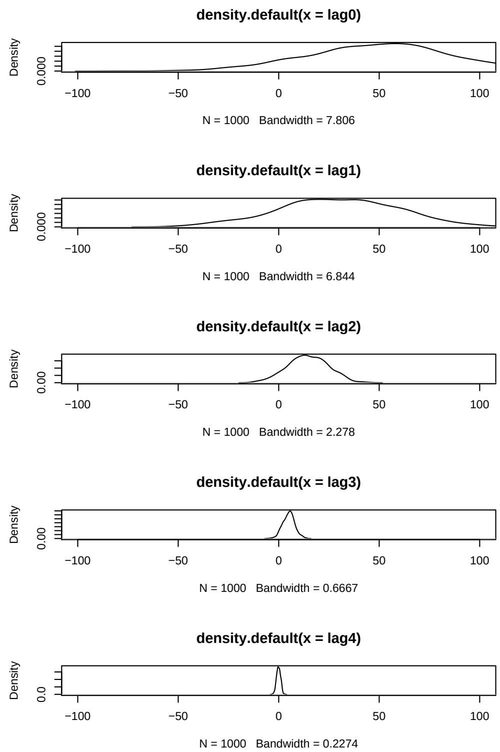
Sims-Zha prior (Princeton) is actually a set of priors that you can bend and shape to your beliefs through varying a set of hyperparameters. These are:

- $\lambda_0, [0, 1]$: Scale of error covariance matrix (1 would mean that prior=sample data).
- $\lambda_1, > 0$: Stand. deviation about A_1 (persistence).
- $\lambda_2, = 1$: Weight of own lag versus other lags
- $\lambda_3, > 0$: Lag decay (quick or slow).
- $\lambda_4, \geq 0$: Scale of standard deviation of intercept (related to drift and equilibrium).
- $\lambda_5, \geq 0$: Scale of standard deviation of exogenous coefficients.
- $\mu_5, \geq 0$: Sum of own AR coefficients (nonstationarity, ∞ implies first differences for all variables).
- $\mu_6, \geq 0$: Common stochastic trends (cointegration).
- $\nu, > 0$: Prior degrees of freedom (m+1)

Sims-Zha Prior example

```
 lag0<-rnorm(1000,50,35); lag1<-rnorm(1000,30,30);
lag2<-rnorm(1000,15,10); lag3<-rnorm(1000,5,3);
lag4<-rnorm(1000,0,1); par(mfrow=c(5,1));
plot(density(lag0), xlim=c(-100,100))
plot(density(lag1), xlim=c(-100,100))
plot(density(lag2), xlim=c(-100,100))
plot(density(lag3), xlim=c(-100,100))
plot(density(lag4), xlim=c(-100,100))
```

Minnesota Prior



Nonstationary Data with B-SVAR

Three setting decide your prior on non-stationarity of data:

1. λ_1: set it tight if you are sure there are unit roots (tightness with stationarity means you think the shocks die out fast.).
2. μ_5: the larger the value to greater the non-stationarity (sum of the coefficients).
3. μ_6: the larger the value the greater the number of cointegrating relationships.

Non-stationarity

Benefits of Bayesian thinking on Non-stationarity:

- Bayesians like John Williams and Patrick Brandt would argue that non-stationarity is a prior belief, since finite sample data can not tell you much about the infinite horizon.
- DF, ADF and KPSS tests all just use sample data to see if they are consistent with re-equilibriation or not. But all use finite sample data.
- In B-SVAR set up with Sims-Zha prior, data can pull your stationary priors towards non-stationary (in IRFs, you will see that impulses have long-lasting effects), or vice-versa.

Error Bands

- What is our uncertainty around impulse response functions? Remember: IRFs are nonlinear transformations of the coefficients in the underlying SVAR or VAR.
  - Serially correlated: Uncertain today, we will be uncertain tomorrow.
  - Non-normal: Non-linear transformations.
  - Asymmetric: We might be more sure that a shock is not way up, rather than it is way down.
  - Standard confidence intervals have been shown to have very poor coverage (too big in some places, too small in many places.)

  Bayesian approach – sample from the posterior (like the election example where you vote vote share and who is going to win. Here we want coefficients and their non-linear moving average representation.)

The MSBVAR Package

- Building on work by Sims and Zha, Patrick Brandt (UT-Dallas) has built an R package that impliments frequentists VARs and B-SVARs.
- Need to set hyper-parameters and then away you go.
- `library(MSVAR)`, also requires xtable, code and lattice packages.
- All available from `www.cran.r-project.org`.

Review Questions

- What is stationarity?
- What are some specific processes that cause non-stationarity?
- How and why do you transform variables towards stationarity?
- What is the equilibrium of a dynamic model? What does that mean?
- ARIMA models: what are they and how are the components related to each other.
- How do we use ACF's and PACF's to identify ARIMA processes?
- What are the uses of ARIMA modeling, what are the hazards?
- What is a jarque-bera test? Explain how it works.
- How do we test for non-stationarity? What are the benefits and the problems with the Dickey-Fuller test and the Augmented Dickey-Fuller test.

Review II

- What are the different types of intervention models?
- What is a transfer function? Why do you have to filter both sides with same filter?
- What is an ADL model, and what are the multipliers and lags of interest?
- What is a Koyck transformation?
- What does it mean mathmatically if two variables are cointegrated?
- What is an error correction model? How do you interpret each component of the model? What is the relationship between an Error correction model and an ADL.
- How would you go about deciding whether two variables are cointegrated?
- What are the pro's and cons of time series regression analysis versus simultaneous equation modeling or static regressions?