Forecasting Economic and Financial Time Series
Week 5: ARDL, VAR's and Basic Bayes for forecasting,

Ron Smith
MSc/PGCE Option, EMS Birkbeck

Spring 2019

Introduction

- We have looked at deterministic models, trend and seasonals, and ARIMA models and now going to look at ARDL and VARs.
- These models are all very closely related and what look like very different representations can give similar degrees of fit and similar forecasts.
- So will remind you of ARMA & ARDL and show how they can be related through the transfer function.
- Purpose matters.
  - Do you want a forecast conditional on information available now (ARIMA, VAR) or do you want to condition on the future path of some exogenous, X, variable (ARDL, VARX).
  - Do you want a forecast for a single variable (ARIMA, ARDL) or for a set of variables (VAR, VARX).
- Parsimonious models tend to forecast better, Bayesian shrinkage estimation is one way to get parsimony.

Alternative ways of estimating AR in EViews

We can write an AR1 model either as

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t; \tag{1}$$

with $E(u_t) = \sigma_\varepsilon^2$

$$E(y_t) = \frac{\alpha}{1-\rho}; \quad V(y_t) = \frac{\sigma^2}{1-\rho^2}$$

or as

$$\begin{aligned} y_t &= \mu + u_t; \quad u_t = \rho u_{t-1} + \varepsilon_t \\ y_t &= \mu + \rho(y_{t-1} - \mu) + \varepsilon_t \\ y_t &= \mu(1-\rho) + \rho y_{t-1} + \varepsilon_t \end{aligned} \tag{2}$$

So if they are both estimated by OLS the only difference is in the value of the intercept: below 4.47=0.70/(1-0.84).

R C R(-1)

Dependent Variable: R
Method: Least Squares
Date: 01/30/17 Time: 13:51
Sample (adjusted): 1872 2011
Included observations: 140 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.704149	0.258426	2.724758	0.0073
R(-1)	0.842630	0.047029	17.91732	0.0000

R-squared	0.699366	Mean dependent var	4.703393
Adjusted R-squared	0.697188	S.D. dependent var	2.800488
S.E. of regression	1.541063	Akaike info criterion	3.717005
Sum squared resid	327.7327	Schwarz criterion	3.759028
Log likelihood	-258.1903	Hannan-Quinn criter.	3.734082
F-statistic	321.0305	Durbin-Watson stat	1.978375
Prob(F-statistic)	0.000000		

R C AR(1)

Dependent Variable: R
Method: ARMA Conditional Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/30/17 Time: 14:35
Sample (adjusted): 1872 2011
Included observations: 140 after adjustments
Convergence achieved after 3 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.474489	0.831601	5.380575	0.0000
AR(1)	0.842630	0.047029	17.91732	0.0000

R-squared	0.699366	Mean dependent var	4.703393
Adjusted R-squared	0.697188	S.D. dependent var	2.800488
S.E. of regression	1.541063	Akaike info criterion	3.717005
Sum squared resid	327.7327	Schwarz criterion	3.759028
Log likelihood	-258.1903	Hannan-Quinn criter.	3.734082
F-statistic	321.0305	Durbin-Watson stat	1.978375
Prob(F-statistic)	0.000000		

Inverted AR Roots	.84

Maximum Likelihood

If (2) is estimated by ML, there is a term in the likelihood: $0.5\log(1-\rho^2)$ that pushes ρ away from unity, so the estimates are different, lower value of ρ.

Dependent Variable: R
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 01/30/17 Time: 13:51
Sample: 1871 2011
Included observations: 141
Convergence achieved after 21 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.621973	0.988455	4.675957	0.0000
AR(1)	0.838465	0.038783	21.61937	0.0000
SIGMASQ	2.331278	0.180304	12.92970	0.0000

R-squared	0.699212	Mean dependent var	4.715071
Adjusted R-squared	0.694853	S.D. dependent var	2.793912
S.E. of regression	1.543359	Akaike info criterion	3.735458
Sum squared resid	328.7103	Schwarz criterion	3.798197
Log likelihood	-260.3498	Hannan-Quinn criter.	3.760953
F-statistic	160.3978	Durbin-Watson stat	1.966208
Prob(F-statistic)	0.000000		

Inverted AR Roots	.84

ARMA(p,q) Model

$$\begin{aligned} D(L)y_t &= C(L)\varepsilon_t \\ y_t &= \frac{C(L)}{D(L)}\varepsilon_t \end{aligned}$$

$$\begin{aligned} D(L) &= (1 - \alpha_1 L - ... - \alpha_p L^p) \\ C(L) &= (1 + \gamma_1 L + ... + \gamma_q L^q) \end{aligned}$$

$$E(y_t) = \alpha_0 / (1 - \sum_{i=1}^{p} \alpha_i)$$

ARMA(1,1)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t + \gamma \varepsilon_{t-1}$$

ARMA Special case of Transfer Function

$$y_t = \frac{A(L)}{B(L)}x_t + \frac{C(L)}{D(L)}\varepsilon_t$$

Diebold table 11.1 gives lots of special cases, e.g. ARMA has no x_t.
Transfer function often estimated by "pre-whitening x_t and y_t with ARIMA models and then looking at the cross-correlogram of their inovations.
We will use **autoregressive distributed lag model (ARDL)** $C(L) = 1$, $D(L) = B(L)$

$$\begin{aligned} y_t &= \frac{A(L)}{B(L)}x_t + \frac{1}{B(L)}\varepsilon_t \\ B(L)y_t &= A(L)x_t + \varepsilon_t \end{aligned}$$

We could approximate an infinite distributed lag, DL, on x_t by a "rational lag" (ratio of $A(L)$ to $B(L)$) with a finite AR on y_t and finite DL on x_t.

Autoregressive Distributed Lag Models

- $ARDL(p_y, p_x)$ for x_{jt}, $j = 1, 2, ..., k$ is given by:

$$y_t = \alpha_0 + \sum_{i=1}^{p_y} \alpha_i y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{p_x} \beta_{ij} x_{j,t-i} + \varepsilon_t$$

- Note here we allow the x_{jt} to enter contemporaneously. Using lagged x_t aids one period ahead forecasting but the problem then arises again at two-stage forecasting and beyond
- Need to forecast the x_{T+h}, but often want to make forecasts conditional e.g. on future policy variables or exogenous variables like oil prices in scenario analysis

Forecasting with an ARDL Model 1

Consider first, a one-period-ahead forecast of y_t in the ARDL(p_y, p_x) model, where we collect all terms α_0, x_t

$$\mu_t = \alpha_0 + \sum_{i=0}^{p_x} \beta_i x_{t-i}$$

- Now the ARDL model is just:

$$y_t = \mu_t + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \varepsilon_t$$

Forecasting with an ARDL Model 2

- Writing out the process for the one-step ahead forecast at $T+1$

$$y_{T+1} = \mu_{T+1} + \alpha_1 y_T + ... + \alpha_p y_{t-p+1} + \varepsilon_{T+1}$$

- Then, conditioning on the full set of information available up to time T and on forecasts of the exogenous variables, the one-period-ahead forecasts of y_t would be:

$$y_{T+1,T} = \mu_{T+1,T} + \alpha_1 y_T + ... + \alpha_p y_{t-p+1}$$

where the expected value of ε_{T+1} is set to zero

- To form a prediction interval, density forecast and therefore to gauge uncertainty we are interested in the variance of the forecast error (where we assume $\mu_{T+1} = \mu_{T+1,T}$)

$$e_{T+1,T} = y_{T+1} - y_{T+1,T} = \varepsilon_{T+1}$$

Forecasting with an ARDL Model 3

- Note here three sources of uncertainty:
  - forecasting μ_t there will be uncertainty about the parameters α_i and β_i from sampling variation,
  - the exogenous variables x_{T+1} will have been forecast, then to the extent that these forecasts are themselves imperfect, another source of error will result
  - the error term ε_t, which in the point forecast we set to zero, but where we seek to quantify this uncertainty through the variance term
- The variance, which takes into account the third source of uncertainty only, and is derived from above where we ignore the variation in $\widehat{\mu}_{T+1,T}$ - that is, assuming that the parameters are known and the exogenous variables are forecasted perfectly is:

$$Var[e_{T+1,T}|x_{T+1}, \alpha_0, \beta, y_{T,...}] = Var[\varepsilon_{T+1}] = \sigma^2$$

- In practice often make judgemental adjustments to the forecast, that is set $\varepsilon_{T+h,T}, = \hat{\varepsilon}^J_{T+h,T}$ rather than zero.

VARMAX

Transfer function was:

$$y_t = \frac{A(L)}{B(L)}x_t + \frac{C(L)}{D(L)}\varepsilon_t$$

Could do it for vectors e.g. for $m \times 1$ vector $\mathbf{y}_t$ and $k \times 1$ vector $\mathbf{x}_t$ with $D(L) = B(L)$, VARMAX is

$$B(L)\mathbf{y}_t = A(L)\mathbf{x}_t + C(L)\varepsilon_t.$$

$B(L) = (I_m - B_1 L - ... - B_p L^p)$,
$A(L) = (A_0 + A_1 L + ... + A_s L^s)$ where A_i are $m \times k$ matrices,
$C(L) = (I_m + C_1 L + ... + C_q L^q)$.
In practice, it is very difficult to separate out the VAR and the MA components so in practice VARs or VARXs are used, though there are Bayesian VARMA estimators.

Vector Autoregressions VARs

- Univariate autoregessions involve one variable - in contrast, a multivariate autoregession - that is a vector autoregression, or **VAR** - involves m variables
- In an m variable vector autoregession of order p, or VAR(p), we estimate m different equations
- In each equation, we regress the relevant left-hand-side variable on p lags of itself and p *lags of every other variable (*trends, seasonals and other exogenous variables may also be included, as long as they are included in all equations)
- Therefore the RHS variables are the same in every equation - p lags of every equation

Advantages of VARs

- Do not need to specify which variables are endogenous or exogenous - all are endogenous
- Allows the value of a variable to depend on more than just its own lags or combinations of white noise terms, so more general than ARMA modelling and will have shorter lags.
- Provided that there are no contemporaneous terms on the right hand side of the equations, can simply use OLS separately on each equation
- Forecasts are often better than "traditional structural" models, which impose incorrect theoretical restrictions.
- Even though the errors are correlated across equations, Seemingly unrelated regressions estimator (SURE) does not add to the efficiency of the estimation procedure since all regressions have identical RHS variables.

Disadvantages of VARs

- VAR's are a-theoretical (as are ARMA models)
- How do you decide the appropriate lag length?
- So many parameters! If we have m equations for m variables and we have p lags of each of the variables in each equation, we have to estimate $m(1 + mp)$ parameters. e.g. $m = 3, p = 3$, parameters $= 30$
- Big danger of "overfitting" to special features of the sample
- Do we need to ensure all components of the VAR are stationary?
- May forecast worse than more parsimonious models.

Estimation

- Since the RHS contain only predetermined variables; the error terms are assumed to be serially uncorrelated with a constant variance; and all equations have identical RHS variables.
- Therefore each equation can be estimated using OLS. Moreover, OLS estimates are consistent and asymptotically efficient.
- Example: $m = 2$, $p = 1$ a two-variable VAR(1)

$$\begin{aligned} y_{1,t} &= a_1 + a_{11} y_{1,t-1} + a_{12} y_{2,t-1} + \varepsilon_{1t} \\ y_{2,t} &= a_2 + a_{21} y_{1,t-1} + a_{22} y_{2,t-1} + \varepsilon_{2t} \end{aligned}$$

- Estimate covariance by $\hat{\sigma}_{12} = \sum_{t=1}^{T} \hat{\varepsilon}_{1t} \hat{\varepsilon}_{2t} / T$
- y_{1t} Granger causal for y_{2t} if it helps predict it, here if $a_{21} \neq 0$

VAR(p)

- More generally we can write a VAR(p) model (a VAR model of order p) as:

$$y_t = a + A_1 y_{t-1} + ... + A_p y_{t-p} + u_t$$

where $y_t = (y_{1t}, ..., y_m)'$ is a $(m \times 1)$ random vector, the A_i are fixed $(m \times m)$ coefficient matrices, $a = (a_1, ..., a_m)'$ is a fixed $(m \times 1)$ vector of intercept terms allowing for the possibility of a non zero mean $E(y_t)$.

- The error term $u_t = (u_{1t}, ..., u_{Nt})'$ is a $m-$dimensional white noise process, where $E(u_t) = 0$, $E(u_t u_t') = \Sigma_u$ and $E(u_t u_s') = 0$ for $s \neq t$

VAR Forecast

- As a simple example, from an estimated VAR(1) model

$$y_t = a + A_1 y_{t-1} + u_t$$

- The h-step-ahead forecast, given by the conditional expectation, can be obtained by solving recursively

$$y^f_{T+h,T} = \hat{a} + \hat{A}_1 y^f_{T+h-1,T}$$

that is:

$$\begin{aligned} y^f_{T+h,T} &= (I_m + \hat{A}_1 + \hat{A}_1^2 + ... + \hat{A}_1^{h-1})a + \hat{A}_1^h y_T \\ &= \sum_{i=0}^{h-1} \hat{A}^i a + \hat{A}_1^h y_T \end{aligned}$$

Direct Projection

The VAR estimates a single model and iterates forward to get the h period ahead forecast. The alternative is to estimate a separate model for each horizon. In the single lag case, one would estimate a sequence of models

$$\begin{aligned} y_t &= b_1 + B_1 y_{t-1} + u_t^1, \\ y_t &= b_2 + B_2 y_{t-2} + u_t^2, \\ &\dots \\ y_t &= b_h + B_h y_{t-h} + u_t^h. \end{aligned}$$

With h period ahead forecast of

$$y_{T+h,T}^f = \hat{b}_h + \hat{B}_h y_T.$$

One would expect $\hat{B}_h$ to differ from $\hat{A}_1^h$. The projection method may be more robust to misspecification.

VAR MSE

- The resulting VAR forecast errors are unbiased:

$$E[y_{T+h} - y_{T+h,T}] = 0$$

with a mean square error or forecast error covariance matrix:

$$\Sigma_y(h) = MSE[y_t(h)] = \sum_{i=0}^{h-1} A_1^i \Sigma_u (A_1^i)'$$

- Biased estimators can have smaller mean squared errors than unbiased ones because of the reduction in variance.
- Can write VAR(p) as VAR(1) by using the companion form, which redefines the dependent variable.

VAR MSE converges to variance of y

- The MSE's are monotonically nondecreasing and, for $h \longrightarrow \infty$, the MSE matrices approach the covariance matrix of y_t

$$\Sigma_y(h) \underset{h\longrightarrow\infty}{\longrightarrow} \Sigma_y$$

$$\Gamma_y(0) = \Sigma_y = \sum_{i=0}^{\infty} \Phi_i \Sigma_u \Phi_i'$$

- If the process mean μ is forecast, the MSE matrix of that predictor is just the covariance matrix Σ_y of y_t
- Therefore the optimal long range forecast, $h \longrightarrow \infty$ is the process mean i.e. the past of the process for forecasts of the distant future contains no information.

VAR: Interval Forecasts

- In order to set up interval forecasts, we make assumptions about the distributions of the y_t or the u_t
- Most common to consider the Gaussian processes where $y_t, y_{T+1}, ..., y_{T+h}$ have a multivariate normal distribution for any t and h
- Equivalently, it may be assumed that u_t is Gaussian, that is, the u_t are multivariate normal, $u_t \sim N(0, \Sigma_u)$ and u_t and u_s are independent for $s \neq t$
- Under these conditions the forecast errors are also normally distributed as linear transformation of normal vectors

$$y_{T+h} - y_{T+h,T} = \sum_{i=0}^{h-1} \Phi_i u_{T+h-i} \sim N(0, \Sigma_y(h))$$

VAR: Interval Forecasts 2

- This implies that the forecast errors of the individual components are normal so that

$$\frac{y_{N,T+h} - y_{N,T+h,T}}{\sigma_N(h)} \sim N(0,1)$$

where $y_{N,T+h}$ is the $N-th$ component of y_{T+h} and $\sigma_N(h)$ is the square root of the $N-th$ diagonal element of $\Sigma_y(h)$

- Hence a $(1-\alpha)100\%$ interval forecast, h-periods ahead, for the $N-th$ component of y_t is

$$y_{N,T+h,T} \pm z_{(\alpha/2)}\sigma_N(h)$$

Frequentist or classical statistics

- Regards probabilities as the limits of relative frequencies as the sample size goes to infinity
- regards parameters as fixed numbers;
- imagines some sampling distribution over lots of hpothetical samples of which the data is just one;
- Uses the Neyman-Pearson hypothesis testing framework

Bayesian Statistics

- Regards probabilities as measuring degrees of belief
- Regards Parameters as random variables
- Uses prior distributions for the parameters based on past experience and uses Bayes rule to provides a systematic way to update beliefs
- Estimation and inference.is done conditional on the observed data not sets of hypothetical samples
- You have an explicit loss function and do not test

Reading

- Gary Koop, *Bayesian Econometrics*, Wiley 2003
- Sharon Bertsch McGrayne, *The theory that would not die*, Yale University Press 2011. Non-technical account of the history of Bayes' rule
- In the literature distinguish committed Bayesians from pragmatic Bayesians who only use it for particular problems where frequentist methods don't work well, e.g. Model Selection, DSGEs and VARS and ad hoc Bayesians who use priors and Bayesian interpretations when doing frequentist statistics.

Bayes Theorem

- Bayes Theorem for continuous variables A and B follows from the definitions of conditional probability in terms of the joint and marginal probabilities

$$f(A \mid B) = \frac{f(A,B)}{f(B)}$$

$$f(B \mid A) = \frac{f(A,B)}{f(A)}$$

so

$$f(A,B) = f(B \mid A) f(A)$$

giving Bayes Theorem

$$f(A \mid B) = \frac{f(B \mid A) f(A)}{f(B)}$$

Bayes Rule

- Treat A as the parameter θ and B as the data Y.

$$f(\theta \mid Y) = \frac{f(Y \mid \theta) f(\theta)}{f(Y)}$$

- For data Y and random parameter θ, Bayesian statistics derives the posterior distribution, $f(\theta \mid Y)$, as proportional to the product of the likelihood, $f(Y \mid \theta)$ and the prior distribution, $f(\theta)$, treating $f(Y)$ as a constant (it has no information about θ, so can be ignored).

$$f(\theta \mid Y) \propto f(Y \mid \theta) f(\theta).$$

  We will usually write the $\propto$ as $=$.

  - Need priors, $f(\theta)$. Can use uninformative priors but they are likely to be improper: $f(\theta)$ does not integrate to one.
  - Need to calculate functions of $f(Y \mid \theta) f(\theta)$, e.g. mean. This usually involves integration, now done numerically through Markov Chain Monte Carlo, MCMC, methods making Bayesian methods easier to apply.

Bayesian Estimation

- Bayes rule gives us a posterior distribution for the parameter conditional on the data. Choice of an estimator is a decision problem.
- To choose an estimator we need a loss function. Quadratic loss function gives mean, absolute loss function gives median.
- Suppose that we choose the mean, this is

$$E(\theta) = \int \theta f(\theta \mid Y) d(\theta)$$

  which involves integration, over the support of θ. Similarly, estimating the posterior variance to get a standard error involves integration.
- These integrals can rarely be worked out analytically, an exception is the simple regression case.

Regression

- For standard linear regression model

$$y \sim N(X\beta, \sigma^2 I) = f(y \mid X\beta, \sigma^2 I)$$

- Bayes Rule is

$$f(\beta, \sigma^2 \mid y, X) = \frac{f(y, X \mid \beta, \sigma^2) f(\beta, \sigma^2)}{f(y, X)}$$

- If X is distributed independently of β, σ^2 we can condition on it

$$\begin{aligned} f(\beta, \sigma^2 \quad \mid \quad y, X) &= \frac{f(y \mid X\beta, \sigma^2 I) f(X) f(\beta, \sigma^2)}{f(y \mid X) f(X)}, \\ &= \frac{f(y \mid X\beta, \sigma^2 I) f(\beta, \sigma^2)}{f(y \mid X)} \end{aligned}$$

Priors

- Conjugate priors when combined with the likelihood give a posterior with the same form of distribution.
- In much Bayesian analysis it is more convenient to work with the precision, the inverse of the variance, $h = \sigma^{-2}$ or covariance matrix: $H = V^{-1}$.
- Normal gamma prior

$$f(\beta, h) = N(\beta \mid \underline{\beta}, \underline{H}) f_{\gamma}(h \mid \underline{\sigma^2}, \underline{\nu}_{\sigma})$$

with for β prior mean $\underline{\beta}$, prior precision $\underline{H}$ and for σ^2 $\underline{\sigma^2}, \underline{\nu}_{\sigma}$

- Together with the normal likelihood this gives a normal gamma posterior

Posterior Means

- For regression model $y = X\beta + u$ with least squares estimates $\hat{\beta} = (X'X)^{-1}X'y$, $H(\hat{\beta}) = \hat{\sigma}^{-2}(X'X)$ prior mean $\underline{\beta}$, prior precision $\underline{H}$
- The posterior is normally distributed with a mean which is a matrix weighted average

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

  and precision $\overline{H} = (H(\hat{\beta}) + \underline{H})$.
- Elements of a matrix weighted average vector, do not have to lie between the prior and OLS, $\overline{\beta_i}$ need not be between $\underline{\beta}_i$ and $\hat{\beta}_i$.
- As $T \longrightarrow \infty$ $H(\hat{\beta})$ gets larger, while $\underline{H}$ is constant so asymptotically $\overline{\beta}$ goes to the Maximum Likelihood estimator $\hat{\beta}$.

Interpretation

- This form of the Bayesian regression estimator

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

  can be interpreted as

  - a shrinkage estimator (like Ridge Regression or Lasso), shrinking the least squares estimator to $\underline{\beta}$, which could be zero
  - Combining estimates of β from two different samples.

- Can implement Bayesian estimation by generating data from the prior (e.g. a DSGE model) and adding it to the real data.

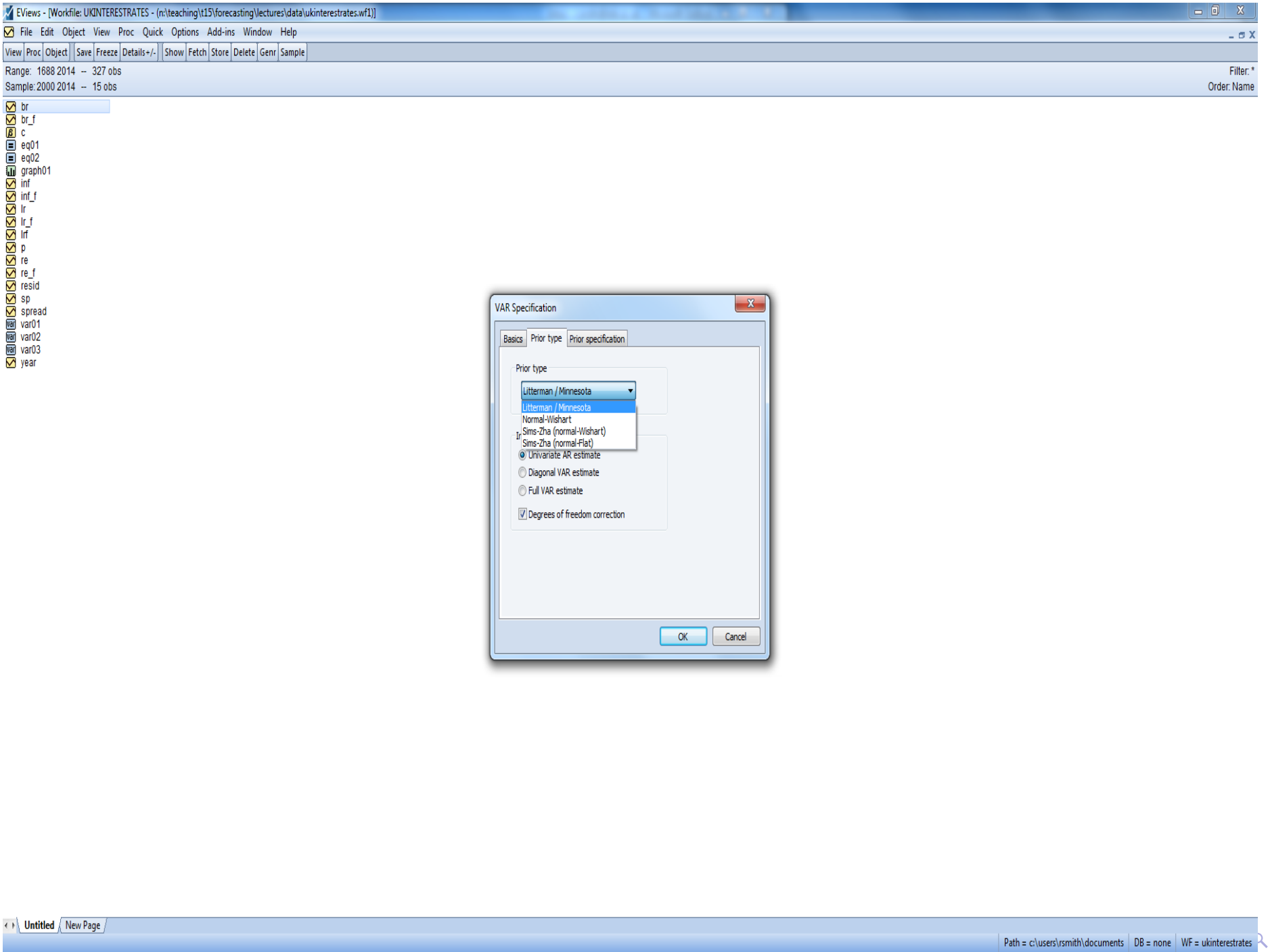
.

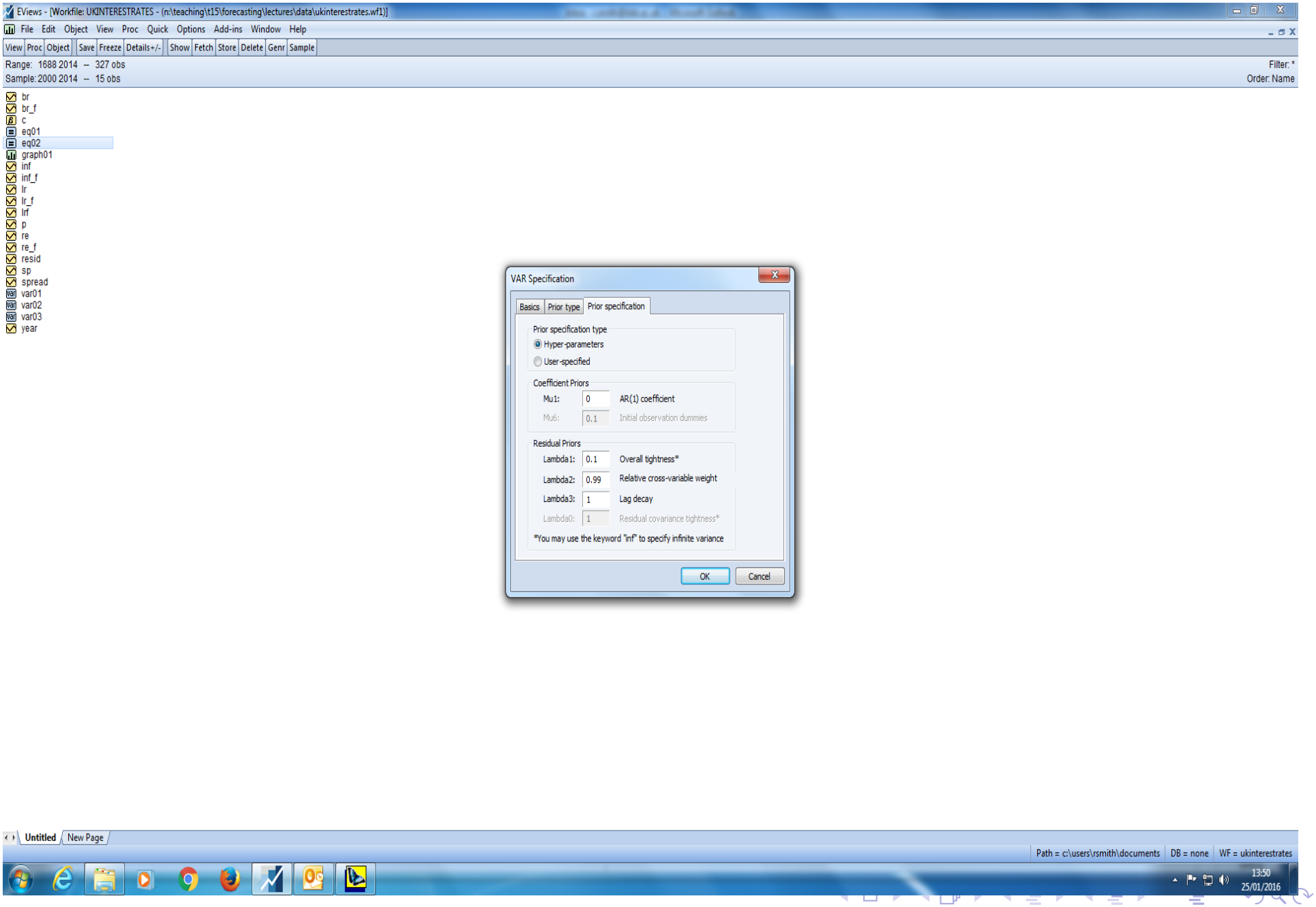
Bayesian VARs

- VARs are just linear regressions so the simple analytical results for Bayesian regression can be applied.
- As noted earlier the Bayesian regression estimator can be regarded as a shrinkage estimator, shrinking the least squares estimator to priors $\underline{\beta}$, which could be zero.
- Priors. Very popular Minnesota (Litterman) Priors, treat all the variables as random walks.
- So put the coefficient of the lagged dependent variable to one if I(1) and all other coefficients to zero.

EViews

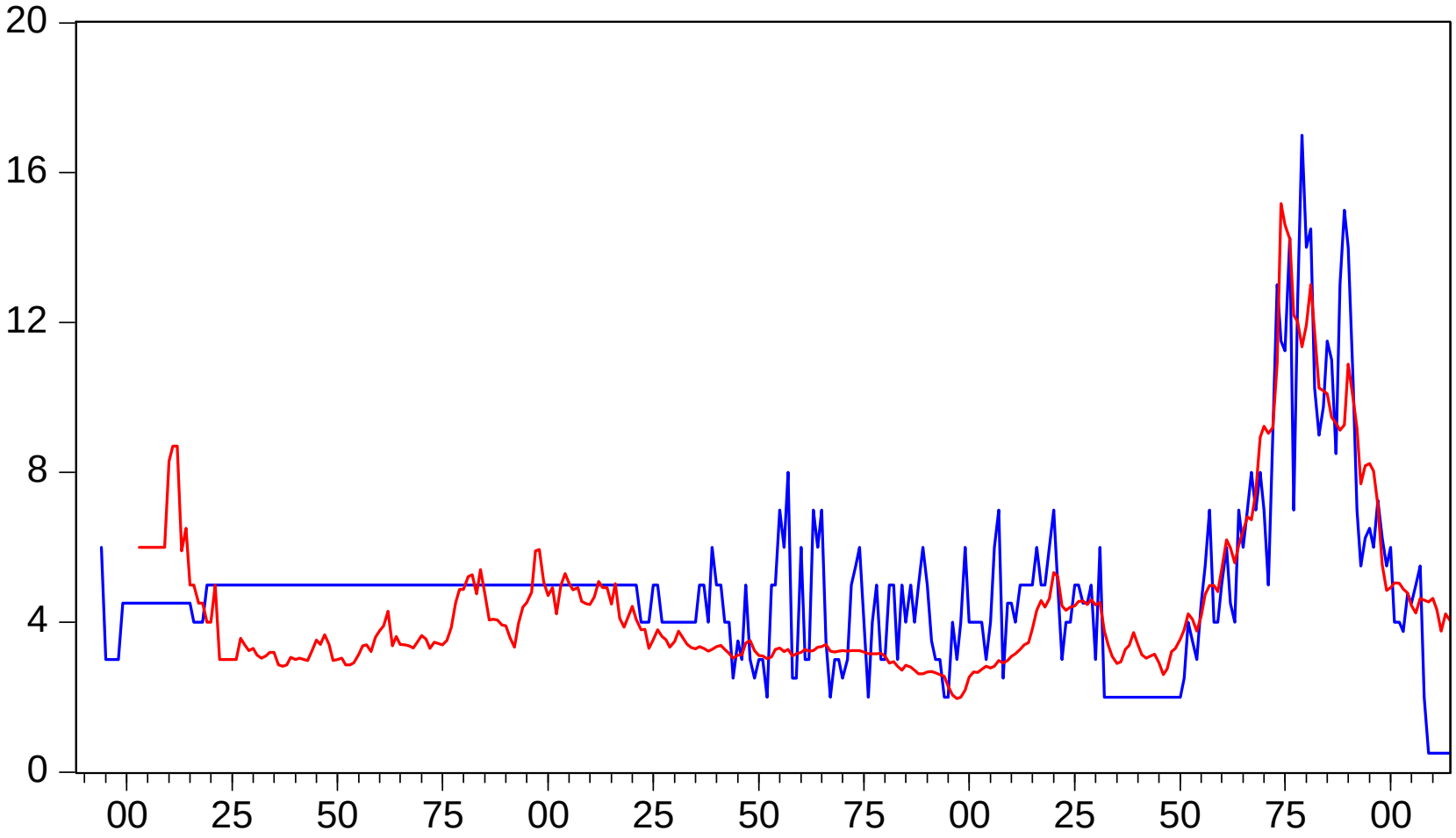
- In EViews Minnesota/Litterman prior
  - Mu1 is the prior for the AR1 coefficient,
  - lambda1 the tightness of the prior for the AR1 coefficient, the smaller lambda1 the tighter the prior, zero imposes the prior,.inf gets you an uniformative prior
  - lambda2 the tightness of the other variables
  - lambda3 the tightness of the lags


EViews - [Workfile: UKINTERESTRATES - (n:\teaching\t15\forecasting\lectures\data\ukinterestrates.wf1)]
File Edit Object View Proc Quick Options Add-ins Window Help
View Proc Object Save Freeze Details+/- Show Fetch Store Delete Genr Sample
Range: 1688 2014 -- 327 obs
Sample: 2000 2014 -- 15 obs
Filter: *
Order: Name
br
br_f
c
eq01
eq02
graph01
inf
inf_f
lr
lr_f
lrf
p
re
re_f
resid
sp
spread
var01
var02
var03
year
VAR Specification
Basics
Prior type
Prior specification
Prior type
Litterman / Minnesota
Litterman / Minnesota
Normal-Wishart
Sims-Zha (normal-Wishart)
Sims-Zha (normal-Flat)
Univariate AR estimate
Diagonal VAR estimate
Full VAR estimate
Degrees of freedom correction
OK
Cancel
Untitled
New Page
Path = c:\users\rsmith\documents
DB = none
WF = ukinterestrates


EViews - [Workfile: UKINTERESTRATES - (n:\teaching\t15\forecasting\lectures\data\ukinterestrates.wf1)]
File Edit Object View Proc Quick Options Add-ins Window Help
View Proc Object Save Freeze Details+/- Show Fetch Store Delete Genr Sample
Range: 1688 2014 -- 327 obs
Sample: 2000 2014 -- 15 obs
Filter: *
Order: Name
br
br_f
c
eq01
eq02
graph01
inf
inf_f
lr
lr_f
lrf
p
re
re_f
resid
sp
spread
var01
var02
var03
year
VAR Specification
Basics
Prior type
Prior specification
Prior specification type
Hyper-parameters
User-specified
Coefficient Priors
Mu1: 0 AR(1) coefficient
Mu6: 0.1 Initial observation dummies
Residual Priors
Lambda1: 0.1 Overall tightness*
Lambda2: 0.99 Relative cross-variable weight
Lambda3: 1 Lag decay
Lambda0: 1 Residual covariance tightness*
*You may use the keyword "inf" to specify infinite variance
OK
Cancel
Untitled
New Page
Path = c:\users\rsmith\documents
DB = none
WF = ukinterestrates
13:50
25/01/2016

Example UK interest rates and inflation

- br bank rate, 1694-2014
- lr long rate (gilts) 1703-2014
- p price index 1688-2014, inf inflation, (% change)
- sp stock price index 1700-2014, re return (% change)
- Source:Bank of England, A Millennium of Data


20
16
12
8
4
0
00
25
50
75
00
25
50
75
00
25
50
75
00

Issues in forecasting of long rates using a VAR

- How many variables to include in the VAR? lr br inf re
- Transformations of variables? e.g. differences of logs, inf re.
- How long a sample? When to start? 1950
- Observations to save for ex post forecasts: 2011-2014
- Lag lengths?
- Allow for cointegration? At 5% trace & max eigenvalue tests indicate 1 CV and re is I(0).
- Frequentist or Bayesian VAR

VAR Lag Order Selection Criteria
Endogenous variables: BR LR RE INF
Exogenous variables: C
Date: 11/25/15 Time: 14:50
Sample: 1950 2010
Included observations: 61

Lag	LogL	LR	FPE	AIC	SC	
0	-691.7966	NA	94987.95	22.81300	22.95142	2
1	-570.8467	222.0719	3047.232	19.37202	20.06411*	1
2	-546.1591	42.09025	2308.504	19.08719	20.33295	1
3	-520.3672	40.59058	1704.741	18.76614	20.56557	1
4	-499.5628	30.01295	1506.699	18.60862	20.96172	1
5	-477.6742	28.70633	1314.617	18.41555	21.32233	1
6	-459.6723	21.24817	1343.493	18.34991	21.81036	1
7	-430.5645	30.53938*	993.9174*	17.92015	21.93427	1
8	-413.1458	15.99088	1139.810	17.87363*	22.44143	1

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Vector Autoregression Estimates
Date: 11/25/15 Time: 14:47
Sample: 1950 2010
Included observations: 61
Standard errors in ( ) & t-statistics in [ ]

	BR	LR	RE	INF
BR(-1)	0.608261	0.110851	-0.882840	0.373007
	(0.14488)	(0.06097)	(1.05505)	(0.15208)
	[4.19852]	[1.81825]	[-0.83677]	[2.45277]
BR(-2)	-0.266212	-0.073109	0.761466	-0.554971
	(0.14738)	(0.06202)	(1.07330)	(0.15471)
	[-1.80629]	[-1.17879]	[0.70946]	[-3.58727]
LR(-1)	0.308158	1.100453	0.085510	1.688352
	(0.40704)	(0.17129)	(2.96430)	(0.42728)
	[0.75706]	[6.42448]	[0.02885]	[3.95144]
LR(-2)	0.324509	-0.223264	1.967461	-1.162404
	(0.37698)	(0.15864)	(2.74535)	(0.39572)
	[0.86081]	[-1.40737]	[0.71665]	[-2.93747]
RE(-1)	0.045024	0.012310	0.071225	-0.009898
	(0.02013)	(0.00847)	(0.14659)	(0.02113)
	[2.23672]	[1.45325]	[0.48586]	[-0.46843]
RE(-2)	0.007131	0.004410	-0.316448	-0.002247
	(0.01918)	(0.00807)	(0.13964)	(0.02013)
	[0.37190]	[0.54650]	[-2.26611]	[-0.11164]
INF(-1)	0.203347	0.104843	-1.758520	0.697935
	(0.11363)	(0.04782)	(0.82748)	(0.11927)
	[1.78962]	[2.19265]	[-2.12515]	[5.85156]
INF(-2)	-0.242688	-0.094139	1.291972	-0.094932
	(0.10475)	(0.04408)	(0.76284)	(0.10996)
	[-2.31684]	[-2.13563]	[1.69363]	[-0.86336]
C	-0.198970	0.490976	-1.892021	-0.453656
	(0.72612)	(0.30556)	(5.28796)	(0.76221)
	[-0.27402]	[1.60679]	[-0.35780]	[-0.59519]
R-squared	0.778738	0.939801	0.266896	0.837730
Adj. R-squared	0.744698	0.930540	0.154111	0.812766
Sum sq. resids	195.7050	34.65671	10379.21	215.6429
S.E. equation	1.939989	0.816379	14.12799	2.036413
F-statistic	22.87698	101.4754	2.366414	33.55676
Log likelihood	-122.1102	-69.31108	-243.2242	-125.0691
Akaike AIC	4.298694	2.567577	8.269645	4.395709
Schwarz SC	4.610134	2.879017	8.581086	4.707150

Granger causality tests for other variables on LR

Dependent variable: LR

Excluded	Chi-sq	df	Prob.
BR	3.540462	2	0.1703
RE	2.746997	2	0.2532
INF	5.726307	2	0.0571
All	12.76738	6	0.0469

Forecast Evaluation
Date: 11/25/15 Time: 15:09
Sample: 2010 2014
Included observations: 5

Variable	Inc. obs.	RMSE	MAE	MA
BR	5	3.540996	3.278040	82.7
LR	5	1.439489	1.213580	42.3
RE	5	8.889960	6.388804	144.
INF	5	6.289948	5.585394	163.

RMSE: Root Mean Square Error
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error
Theil: Theil inequality coefficient

VAR forecast rising long rates

AR1 beats VAR on RMSE and MAE

VAR: RMSE= 1.439, MAE 1.21
AR1: RMSE=0.945, MAE=0.827

Bayesian VAR

- Take long rate and bank rate and VAR1 1950-2010 to see the effects of the priors
- Estimate unrestricted VAR
- Then a Bayesian VAR In EViews Litterman prior
  - Mu1 is the prior for the AR1 coefficient, first set to zero and then to one.
  - lambda1 the tightness of the prior for the AR1 coefficient, the smaller lambda1 the tighter the prior, zero imposes the prior,.inf gets you an uniformative prior=0.1
  - lambda2 the tightness of the other variables=0.99
  - lambda3 the tightness of the lags=1

Unrestricted VAR1

	LR	BR
LR(-1)	0.837616	0.395820
	(0.06700)	(0.16080)
	[12.5012]	[2.46162]
BR(-1)	0.109917	0.561223
	(0.05507)	(0.13216)
	[1.99595]	[4.24652]
C	0.450742	0.171632
	(0.29506)	(0.70810)
	[1.52763]	[0.24238]

Mu=0, lambda1=0.1

	LR	BR
LR(-1)	0.729438	0.615025
	(0.04474)	(0.10920)
	[16.3033]	[5.63217]
BR(-1)	0.118740	0.264624
	(0.03140)	(0.07702)
	[3.78098]	[3.43598]
C	1.204371	0.669824
	(0.28563)	(0.69622)
	[4.21654]	[0.96209]

Mu=1, lambda1=0.1

	LR	BR
LR(-1)	0.929621	0.080280
	(0.04474)	(0.10920)
	[20.7775]	[0.73517]
BR(-1)	0.029608	0.857761
	(0.03140)	(0.07702)
	[0.94280]	[11.1375]
C	0.338813	0.402123
	(0.28563)	(0.69622)
	[1.18619]	[0.57758]

Mu=1, lambda1=0.05

	LR	BR
LR(-1)	0.963199	-0.001250
	(0.03205)	(0.07795)
	[30.0560]	[-0.01604]
BR(-1)	0.007068	0.948700
	(0.01821)	(0.04476)
	[0.38816]	[21.1968]
C	0.248774	0.357627
	(0.25496)	(0.62071)
	[0.97572]	[0.57616]

Warning

- Do not believe that because VARs etc are more complicated they will forecast better. Most of the evidence is that parsimonious (sophistically simple, with few estimated parameters) often forecast better.
  - Daily working day financial data data, random walk $\hat{y}_{T+1} = y_T$, $\hat{y}_{T+2} = y_T$, ..
  - Daily traffic data: $\hat{y}_{T+1} = y_{T-6} + (y_T - y_{T-7})$, $\hat{y}_{T+2} = y_{T-5} + (y_T - y_{T-7})$
- Clements & Hendry, Forecasting Economic Time Series, (CUP, 1998) Chapter 12 on Parsimony discusses this.
  - Robust against structural breaks, e.g. first difference models avoid return to mean after a break
  - Avoids, overfitting to specific features of the estimation sample, which can produce spurious significance
  - Reduces estimation uncertainty.
- How do we know: simple models do well? Discuss with forecast evaluation later..

ARDL Forecasting Example Interest rates

- Dependent variable, long rates, LR
- Independent variable, bank rate BR, the policy variable.
- Want to estimate an ARDL model
- and to forecast the effect on long-rates of the policy rate going to 0.5% as it did in 2009

Dependent Variable: LR
Method: Least Squares
Date: 01/25/16 Time: 13:04
Sample: 1960 2008
Included observations: 49

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.502507	0.373553	1.345208	0.1856
BR	0.198169	0.057326	3.456863	0.0012
LR(-1)	1.078736	0.141431	7.627318	0.0000
BR(-1)	0.010804	0.071516	0.151064	0.8806
LR(-2)	-0.329813	0.131853	-2.501361	0.0163
BR(-2)	-0.013615	0.065747	-0.207078	0.8369

R-squared	0.926781	Mean dependent var	8.370787
Adjusted R-squared	0.918267	S.D. dependent var	2.949919
S.E. of regression	0.843354	Akaike info criterion	2.611417
Sum squared resid	30.58355	Schwarz criterion	2.843068
Log likelihood	-57.97971	Hannan-Quinn criter.	2.699305
F-statistic	108.8551	Durbin-Watson stat	2.068551
Prob(F-statistic)	0.000000		

Dependent Variable: LR
Method: Least Squares
Date: 01/25/16 Time: 13:08
Sample: 1960 2008
Included observations: 49

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.507324	0.364702	1.391064	0.171
BR	0.202081	0.049020	4.122431	0.000
LR(-1)	1.080656	0.126029	8.574674	0.000
LR(-2)	-0.338829	0.122268	-2.771195	0.008

R-squared	0.926699	Mean dependent var	8.37078
Adjusted R-squared	0.921813	S.D. dependent var	2.94991
S.E. of regression	0.824856	Akaike info criterion	2.53089
Sum squared resid	30.61740	Schwarz criterion	2.68532
Log likelihood	-58.00682	Hannan-Quinn criter.	2.58948
F-statistic	189.6371	Durbin-Watson stat	2.05023
Prob(F-statistic)	0.000000		

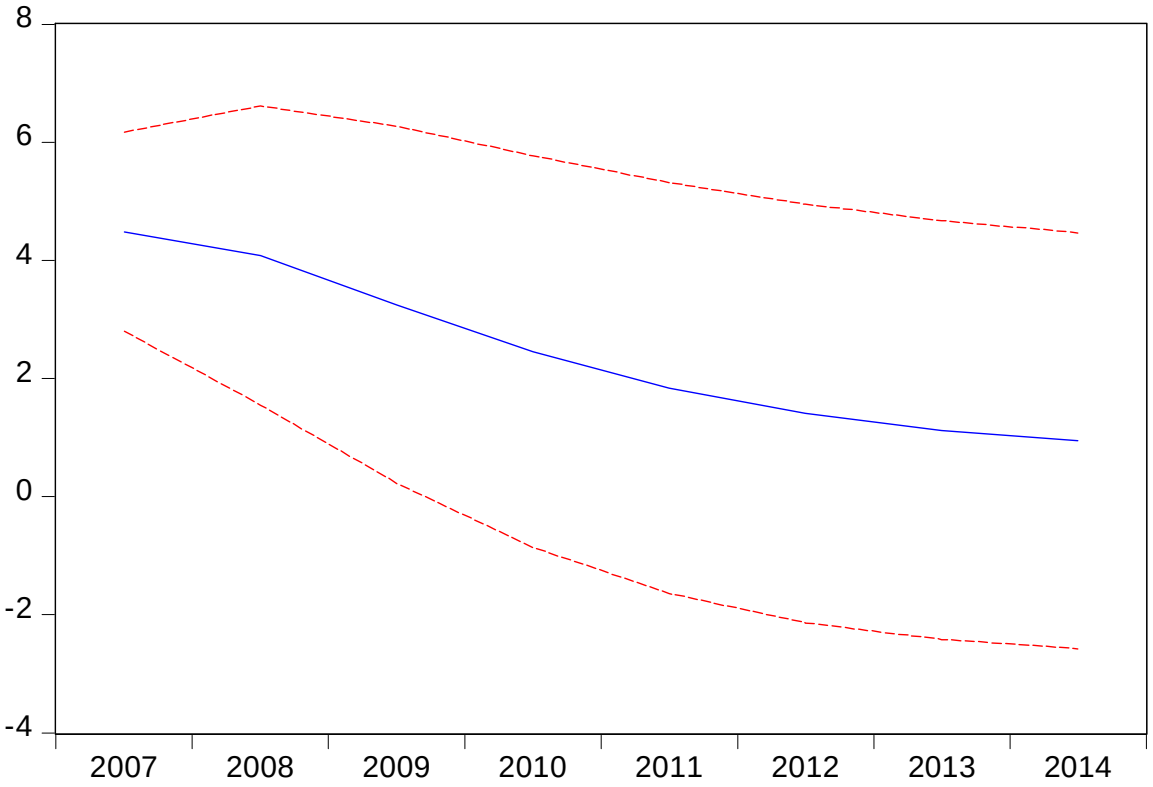
Coefficients suggest a restricted version which has almost identical AIC

Dependent Variable: D(LR)
Method: Least Squares
Date: 01/25/16 Time: 13:13
Sample: 1960 2008
Included observations: 49

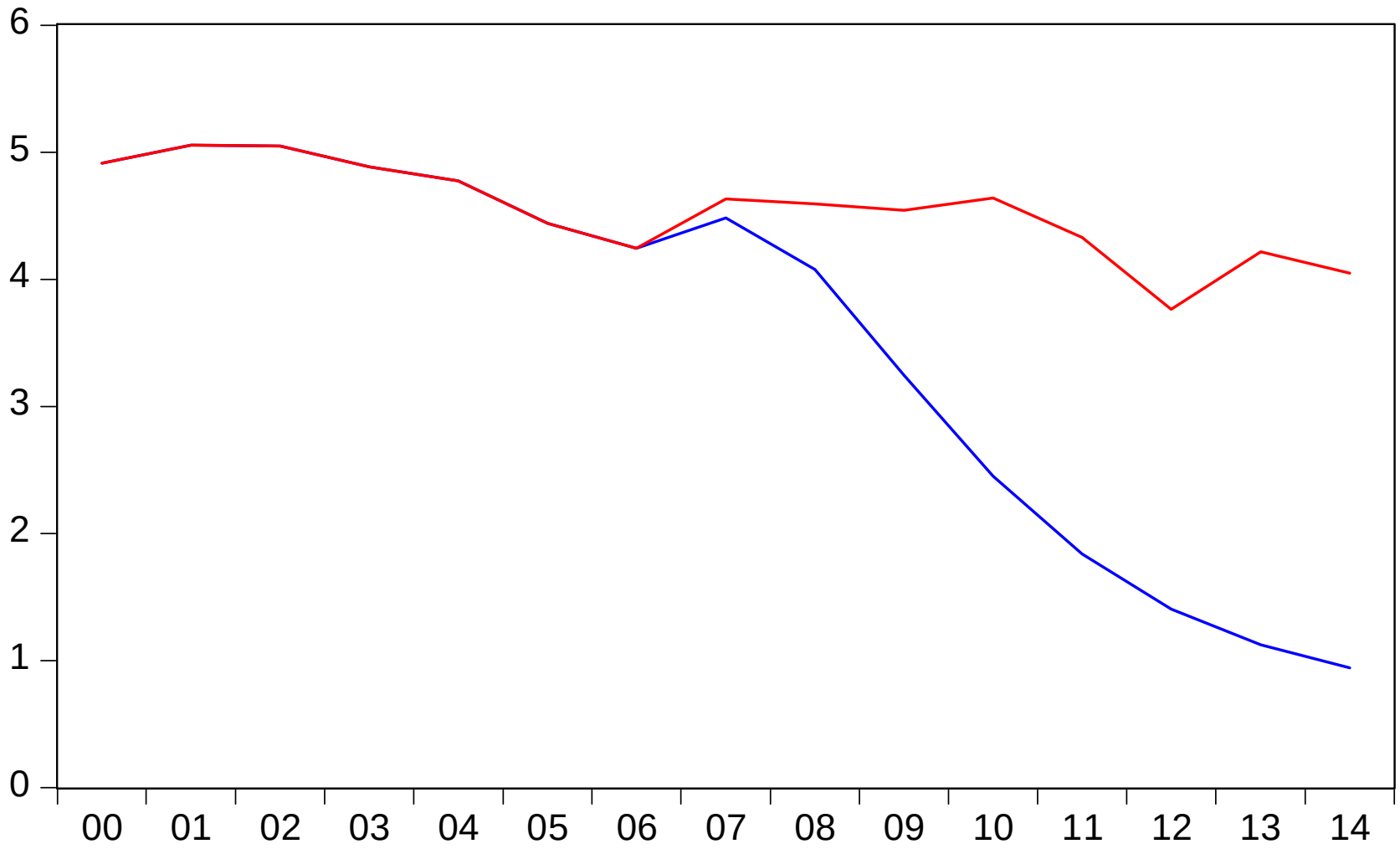
Variable	Coefficient	Std. Error	t-Statistic	Prob
C	0.038591	0.119320	0.323422	0.747
BR-LR(-1)	0.208488	0.049239	4.234185	0.000
D(LR(-1))	0.311046	0.121649	2.556905	0.013

R-squared	0.327115	Mean dependent var	-0.00466
Adjusted R-squared	0.297860	S.D. dependent var	0.99339
S.E. of regression	0.832405	Akaike info criterion	2.53027
Sum squared resid	31.87332	Schwarz criterion	2.64610
Log likelihood	-58.99174	Hannan-Quinn criter.	2.57421
F-statistic	11.18119	Durbin-Watson stat	2.01376
Prob(F-statistic)	0.000110		

Forecast



Forecast too low



Exercises on Lag Operators A: Koyck Lag 1

Desired capital stock, K_t^* depends on income and prices. Neither desired nor actual capital stock is observed but gross investment, change in stock plus depreciation, is observed. Show that capital stock is an infinite depreciated sum of past investment (3). Change in stock is proportional to the difference between desired stock and actual stock in the previous period plus an error.

$$
\begin{aligned}
K_t^* &= \theta_0 + \theta_y y_t + \theta_x p_t \\
I_t &= K_t - K_{t-1} + \delta K_{t-1} = (1 - (1-\delta)L)K_t \\
K_t &= \sum_{i=0}^{\infty} (1-\delta)^i I_{t-i} = (1-(1-\delta)L)^{-1} I_t \qquad (3) \\
\Delta K_t &= \lambda(K_t^* - K_{t-1}) + \varepsilon_t \\
I_t &= \lambda K_t^* + (\delta - \lambda) K_{t-1} + \varepsilon_t
\end{aligned}
$$

Koyck Lag 2

1. Show that the model

$$I_t = \lambda(\theta_0 + \theta_y y_t + \theta_x p_t) + (\delta - \lambda)K_{t-1} + \varepsilon_t$$

which involves an infinite distributed lag on I_t can be written

$$\begin{aligned} I_t &= \delta\lambda\theta_0 + \lambda\theta_y(y_t - (1-\delta)y_{t-1}) + \lambda\theta_p(p_t - (1-\delta)p_{t-1}) \quad (4) \\ &\quad + (1-\lambda)I_{t-1} + \varepsilon_t - (1-\delta)\varepsilon_{t-1}. \end{aligned}$$

2. Show that (4) is over-identifed and explain what the restrictions are
3. Discuss how you would estimate (4) and test the restrictions.
4. Explain how you would get an estimate of capital stock.
5. How else might you estimate the model.

Exercises on Lag Operators B: VARMAX

For $m \times 1$ vector $\mathbf{y}_t$ and $k \times 1$ vector $\mathbf{x}_t$ with $D(L) = B(L)$, VARMAX is

$$B(L)\mathbf{y}_t = A(L)\mathbf{x}_t + C(L)\varepsilon_t$$

Suppose $m = 2$, $k = 1$, write out the two equations of a VARMAX(1,1,1), where the lag order on all three components is one.
Writing it out first in matrix form, with the dimensions of the matrices might help.
How many parameters do you need to estimate?

Answer

$B(L) = (I_m - B_1 L)$,
$A(L) = (A_0 + A_1 L)$ where A_i are $m \times k$ matrices,
$C(L) = (I_m + C_1 L)$

$$y_t = B_1 y_{t-1} + A_0 x_t + A_1 x_{t-1} + \varepsilon_t + C_1 \varepsilon_t$$

$$\begin{aligned} y_{1t} &= b_{11} y_{1,t-1} + b_{12} y_{2,t-1} + a_{10} x_t + a_{11} x_{t-1} + \varepsilon_{1t} + c_{11} \varepsilon_{1,t-1} + c_{12} \varepsilon_{2,t-1} \\ y_{2t} &= b_{21} y_{1,t-1} + b_{22} y_{2,t-1} + a_{20} x_t + a_{21} x_{t-1} + \varepsilon_{1t} + c_{21} \varepsilon_{1,t-1} + c_{22} \varepsilon_{2,t-1} \end{aligned}$$