MA Advanced Macroeconomics
2. Vector Autoregressions

Karl Whelan

School of Economics, UCD

Spring 2016

Part I

Introducing VAR Methods

Background on VARs

- These model were introduced to the economics profession by Christopher Sims (1980) in a path-breaking article titled "Macroeconomics and Reality."
- Sims was indeed telling the macro profession to "get real."
- He criticized the widespread use of highly specified macro-models that made very strong identifying restrictions (in the sense that each equation in the model usually excluded most of the model's other variables from the right-hand-side) as well as very strong assumptions about the dynamic nature of these relationships.
- VARs were an alternative that allowed one to model macroeconomic data accurately, without having to impose lots of incredible restrictions. In the phrase used in an earlier paper by Sargent and Sims (who shared the Nobel prize award) it was "macro modelling without pretending to have too much a priori theory."
- We will see that VARs are not theory free. But they do make the role of theoretical identifying assumptions far clearer than was the case for the types of models Sims was criticizing.

Matrix Formulation of VARs

- The simplest possible VAR features two variables and one lag:

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

- The most compact way to express a VAR system like this is to use matrices. Defining the matrices

$$\begin{aligned} Y_t &= \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} \\ A &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \\ e_t &= \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \end{aligned}$$

- This system can be written as

$$Y_t = AY_{t-1} + e_t$$

Vector Moving Average (VMA) Representation

- VARs express variables as function of what happened yesterday and today's shocks.
- But what happened yesterday depended on yesterday's shocks and on what happened the day before.
- This VMA representation is obtained as follows

$$\begin{aligned} Y_t &= e_t + AY_{t-1} \\ &= e_t + A\left(e_{t-1} + AY_{t-2}\right) \\ &= e_t + Ae_{t-1} + A^2\left(e_{t-2} + AY_{t-3}\right) \\ &= e_t + Ae_{t-1} + A^2 e_{t-2} + A^3 e_{t-3} + ... + A^t e_0 \end{aligned}$$

- This makes clear how today's values for the series are the cumulation of the effects of all the shocks from the past.
- It is also useful for deriving predictions about the properties of VARs.

Impulse Response Functions

- Suppose there is an initial shock defined as

$$e_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

  and then all error terms are zero afterwards, i.e. $e_t = 0$ for $t > 0$.

- Recall VMA representation

$$Y_t = e_t + Ae_{t-1} + A^2 e_{t-2} + A^3 e_{t-3} + ... + A^t e_0$$

- This tells us that the response after n periods is $A^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

- So IRFs for VARs are directly analagous to the IRFs for AR(1) models that we looked at before.

Using a VAR to Forecast

- VARs are often used for forecasting.
- Supppose we observe our vector of variables Y_t. What's our forecast for Y_{t+1}?
- The model for next period is

$$Y_{t+1} = AY_t + e_{t+1}$$

- Because $E_t e_{t+1} = 0$, an unbiased forecast at time t is AY_t. In other words, $E_t Y_{t+1} = AY_t$.
- The same reasoning tells us that $A^2 Y_t$ is an unbiased forecast of Y_{t+2} and $A^3 Y_t$ is an unbiased forecast of Y_{t+3} and so on.
- So once a VAR is estimated and organised to be in this form, it is very easy to construct forecasts.

Generality of the First-Order Matrix Formulation: I

- The model we've been looking at may seem like a small subset of all possible VARs because it doesn't have a constant term and only has lagged values from one period ago.
- However, one can add a third variable here which takes the constant value 1 each period. The equation for the constant term will just state that it equals its own lagged values. So this formulation actually incorporates models with constant terms.
- We would also expect most equations in a VAR to have more than one lag. Surely this makes things much more complicated?
- Not really. It turns out, the first-order matrix formulation can represent VARs with longer lags.
- Consider the two-lag system

$$\begin{aligned} y_{1t} &= a_{11}y_{1,t-1} + a_{12}y_{1,t-2} + a_{13}y_{2,t-1} + a_{14}y_{2,t-2} + e_{1t} \\ y_{2t} &= a_{21}y_{1,t-1} + a_{22}y_{1,t-2} + a_{23}y_{2,t-1} + a_{24}y_{2,t-2} + e_{2t} \end{aligned}$$

Generality of the First-Order Matrix Formulation: II

- Now define the vector

$$Z_t = \begin{pmatrix} y_{1t} \\ y_{1,t-1} \\ y_{2t} \\ y_{2,t-1} \end{pmatrix}$$

- This system can be represented in matrix form as

$$Z_t = AZ_{t-1} + e_t$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 1 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad e_t = \begin{pmatrix} e_{1t} \\ 0 \\ e_{2t} \\ 0 \end{pmatrix}$$

- This is sometimes called the "companion form" matrix formulation.

Interpreting Shocks and Impulse Responses

- The system we've been looking at is usually called a *reduced-form* VAR model.
- It is a purely *econometric* model, without any theoretical element.
- How should we interpret it? One interpretation is that e_{1t} is a shock that affects only y_{1t} on impact and e_{2t} is a shock that affects only y_{2t} on impact.
- For instance, one can use the IRFs generated from an inflation-output VAR to calculate the dynamic effects of "a shock to inflation" and "a shock to output".
- But other interpretations are available.
- For instance, one might imagine that the true shocks generating inflation and output are an "aggregate supply" shock and an "aggregate demand" shock and that both of these shocks have a direct effect on both inflation and output.
- How would we identify these "structural" shocks and their impulse responses?

The Multiplicity of Shocks and IRFs

- Suppose reduced-form and structural shocks are related by

$$\begin{aligned} e_{1t} &= c_{11}\epsilon_{1t} + c_{12}\epsilon_{2t} \\ e_{2t} &= c_{21}\epsilon_{1t} + c_{22}\epsilon_{2t} \end{aligned}$$

- Can write this in matrix form as

$$e_t = C\epsilon_t$$

- These two VMA representations describe the data equally well:

$$\begin{aligned} Y_t &= e_t + Ae_{t-1} + A^2e_{t-2} + A^3e_{t-3} + ... + A^te_0 \\ &= C\epsilon_t + AC\epsilon_{t-1} + A^2C\epsilon_{t-2} + A^3C\epsilon_{t-3} + ... + A^tC\epsilon_0 \end{aligned}$$

- Can interpret the model as one with shocks e_t and IRFs given by A^n.
- Or as a model with structural shocks ϵ_t and IRFs are given by A^nC.
- And we could do this for any C: We just don't know the structural shocks.

Contemporaneous Interactions: I

- Another way to see how reduced-form shocks can be different from structural shocks is if there are contemporaneous interactions between variables, which is likely.
- Consider the following model:

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

- Can be written in matrix form as

$$AY_t = BY_{t-1} + \epsilon_t$$

where

$$A = \begin{pmatrix} 1 & -a_{12} \\ -a_{21} & 1 \end{pmatrix}$$

Contemporaneous Interactions: II

- Now if we estimate the "reduced-form" VAR model

$$Y_t = DY_{t-1} + e_t$$

- Then the reduced-form shocks and coefficients are

$$\begin{aligned} D &= A^{-1}B \\ e_t &= A^{-1}\epsilon_t \end{aligned}$$

- Again, the following two decompositions both describe the data equally well

$$\begin{aligned} Y_t &= e_t + De_{t-1} + D^2e_{t-2} + D^3e_{t-3} + ... \\ &= A^{-1}\epsilon_t + DA^{-1}\epsilon_{t-1} + D^2A^{-1}\epsilon_{t-2} + ... + D^tA^{-1}\epsilon_0 \end{aligned}$$

- For the structural model, the impulse responses to the structural shocks from n periods are given by D^nA^{-1}.
- Again, this is true for any arbitrary A matrix.

Why Care?

- There is no problem with forecasting with reduced-form VARs: Once you know the reduced-form shocks and how they have affected today's value of the variables, you can use the reduced-form coefficients to forecast.
- The problem comes when you start asking "what if" questions? For example, "what happens if there is a shock to the first variable in the VAR?"
- In practice, the error series in reduced-form VARs are usually correlated with each other. So are you asking "What happens when there is a shock to the first variable only?" or are you asking "What usually happens when there is a shock to the first variable given that this is usually associated with a corresponding shock to the second variable?"
- Most likely, the really interesting questions about the structure of the economy relate to the impact of different types of shocks that are uncorrelated with each other.
- A structural identification that explains how the reduced-form shocks are actually combinations of uncorrelated structural shocks is far more likely to give clear and interesting answers.

Structural VARs: A General Formulation

- In its general formulation, the structural VAR is

$$AY_t = BY_{t-1} + C\epsilon_t$$

- The model is fully described by the following parameters:
  1. n^2 parameters in A
  2. n^2 parameters in B
  3. n^2 parameters in C
  4. $\frac{n(n+1)}{2}$ parameters in Σ, which describes the pattern of variances in covariances underlying the shock terms.
- Adding all these together, we see that the most general form of the structural VAR is a model with $3n^2 + \frac{n(n+1)}{2}$ parameters.

Identification of Structural VARs: The General Problem

- Estimating the reduced-form VAR

$$Y_t = DY_{t-1} + e_t$$

gives us information on $n^2 + \frac{n(n+1)}{2}$ parameters: The coefficients in D and the estimated covariance matrix of the reduced-form errors.

- To obtain information about structural shocks, we thus need to impose $2n^2$ *a priori* theoretical restrictions on our structural VAR.
- This will leave us with $n^2 + \frac{n(n+1)}{2}$ known reduced-form parameters and $n^2 + \frac{n(n+1)}{2}$ structural parameters that we want to know.
- This can be expressed as $n^2 + \frac{n(n+1)}{2}$ equations in $n^2 + \frac{n(n+1)}{2}$ unknowns, so we can get a unique solution.
- Example: Asserting that the reduced-form VAR is the structural model is the same as imposing the $2n^2$ a priori restrictions that $A = C = I$.

Recursive SVARs

- SVARs generally identify their shocks as coming from distinct independent sources and thus assume that they are uncorrelated.
- The error series in reduced-form VARs are usually correlated with each other. One way to view these correlations is that the reduced-form errors are combinations of a set of statistically independent structural errors.
- The most popular SVAR method is the recursive identification method.
- This method (used in the original Sims paper) uses simple regression techniques to construct a set of uncorrelated structural shocks directly from the reduced-form shocks.
- This method sets $A = I$ and constructs a C matrix so that the structural shocks will be uncorrelated.

The Cholesky Decomposition

- Start with a reduced-form VAR with three variables and errors e_{1t}, e_{2t}, e_{3t}.
- Take one of the variables and assert that this is the first structural shock, $\epsilon_{1t} = e_{1t}$.
- Then run the following two OLS regressions involving the reduced-form shocks

$$\begin{aligned} e_{2t} &= c_{21} e_{1t} + \epsilon_{2t} \\ e_{3t} &= c_{31} e_{1t} + c_{32} e_{2t} + \epsilon_{3t} \end{aligned}$$

- This gives us a matrix equation $Ge_t = \epsilon_t$.
- Inverting G gives us C so that $e_t = C\epsilon_t$. Identification done.
- Remember that error terms in OLS equations are uncorrelated with the right-hand-side variables in the regressions.
- Note now that, by construction, the ϵ_t shocks constructed in this way are uncorrelated with each other.

Interpreting the Cholesky Decomposition

- The method posits a sort of "causal chain" of shocks.
- The first shock affects all of the variables at time t. The second only affects two of them at time t, and the last shock only affects the last variable at time t.
- The reasoning usually relies on arguments such as "certain variables are sticky and don't respond immediately to some shocks." We will discuss examples next week.
- A serious drawback: The causal ordering is not unique. Any one of the VARs variables can be listed first, and any one can be listed last.
- This means there are $n! = (1)(2)(3)....(n)$ possible recursive orderings.
- Which one you like will depend on your own prior thinking about causation.

Another Way to Do Recursive VARs

- The idea of certain shocks having effects on only some variables at time t can be re-stated as some *variables* only having effects on some variables at time t.
- In our 3 equation example this method sets $C = I$ and directly estimates the A and B matrices using OLS:

$$\begin{aligned} y_{1t} &= b_{11}y_{1,t-1} + b_{12}y_{2,t-1} + b_{13}y_{3,t-1} + \epsilon_{1t} \\ y_{2t} &= b_{21}y_{1,t-1} + b_{22}y_{2,t-1} + b_{23}y_{3,t-1} - a_{21}y_{1t} + \epsilon_{2t} \\ y_{3t} &= b_{31}y_{1,t-1} + b_{32}y_{2,t-1} + b_{33}y_{3,t-1} - a_{31}y_{1t} - a_{32}y_{2t} + \epsilon_{3t} \end{aligned}$$

- See how the first shock affects all the variables while the last shock only affects the last variable.
- This method delivers shocks and impulse responses that are identical to the Cholesky decomposition.
- Shows that different combinations of A, B and C can deliver the same structural model.

Part II

Estimating VARs

How to Estimate a VAR's Parameters?

- VARs consist of a set of linear equations, so OLS is an obvious technique for estimating the coefficients.
- However, it turns out that OLS estimates of a VAR's parameters are biased.
- Here we discuss a set of econometric issues relating to VARs. Specifically, we discuss

1. The nature of the bias in estimating VARs with OLS.
2. Methods for adjusting OLS estimates for their bias.
3. An asymptotic justification for OLS, stemming from the fact that they are Maximum Likelihood Estimators for VAR models when errors are normal.
4. A method for calculating standard errors for impulse response functions.
5. Problems due to have large amounts of parameters to estimate.
6. Bayesian estimation of VAR models as a solution to this problem.

OLS Estimates of VAR Models Are Biased

- Consider the AR(1) model

$$y_t = \rho y_{t-1} + \epsilon_t$$

- The OLS estimator for a sample of size T is

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_{t=2}^{T} y_{t-1} y_t}{\sum_{t=2}^{T} y_{t-1}^2} \\
&= \rho + \frac{\sum_{t=2}^{T} y_{t-1} \epsilon_t}{\sum_{t=2}^{T} y_{t-1}^2} \\
&= \rho + \sum_{t=2}^{T} \left(\frac{y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \right) \epsilon_t
\end{aligned}$$

- ϵ_t is independent of y_{t-1}, so $\mathbb{E}(y_{t-1}\epsilon_t) = 0$. However, ϵ_t is **not independent** of the sum $\sum_{t=2}^{T} y_{t-1}^2$. If ρ is positive, then a positive shock to ϵ_t raises current and future values of y_t, all of which are in the sum $\sum_{t=2}^{T} y_{t-1}^2$. This means there is a negative correlation between ϵ_t and $\frac{y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2}$, so $\mathbb{E}\,\hat{\rho} < \rho$.
- This argument generalises to VAR models: OLS estimates are biased.

The Bias of VAR Estimates

- Recall that for the AR(1) model, the OLS estimate can be written as

$$\hat{\rho} = \rho + \sum_{t=2}^{T} \left(\frac{y_{t-1}}{\sum_{t=2}^{T} y_{t-1}^2} \right) \epsilon_t$$

- Bias stems from correlation of ϵ_t with $\sum_{t=2}^{T} y_{t-1}^2$.
- The size of the bias depends on two factors:
  1. The size of ρ: The bigger this is, the stronger the correlation of the shock with future values and thus the bigger the bias.
  2. The sample size T: The larger this is, the smaller the fraction of the observations sample that will be highly correlated with the shock and thus the smaller the bias.
- More generally, the bias in OLS estimates of VAR coefficients will be larger the higher are the "own lag" coefficients and the smaller the sample size.

A Bootstrap Bias Adjustment

- If OLS bias is likely to be a problem, one solution is to use "bootstrap methods".
- These use the estimated error terms to simulate the underlying sampling distribution of the OLS estimators when the data generating process is given by a VAR with the estimated parameters. These calculations can be used to apply an adjustment to the OLS bias.
- In practice, this can be done roughly as follows:
  1. Estimate the VAR $Z_t = AZ_{t-1} + \epsilon_t$ via OLS and save the errors $\hat{\epsilon}_t$.
  2. Randomly sample from these errors to create, for example, 10,000 new error series ϵ_t^* and simulated data series generated by the recursion $Z_t^* = \hat{A}^{OLS} Z_{t-1}^* + \epsilon_t^*$. (Need some starting assumption about Z_0.)
  3. Estimate a VAR model on the simulated data and save the 10,000 different sets of OLS estimate coefficients $\hat{A}^*$.
  4. Compute median values of each entry in $\hat{A}^*$ as $\bar{A}$ and compare this to $\hat{A}$ to get an estimate of the bias of the OLS estimates.
  5. Formulate new estimates $\hat{A}^{BOOT} = \hat{A}^{OLS} - \left(\bar{A} - \hat{A}^{OLS}\right)$
- See the paper by Killian on the webpage.

Maximum Likelihood Estimation

- Let θ be a potential set of parameters for a model and let f be a function such that, when the model parameters equal θ, the joint probability density function that generates the data is given by $f(y_1, y_2, ..., y_n|\theta)$.
- In other words, given a value of θ, $f(y_1, y_2, ..., y_n|\theta)$ describes the probability density of the sample $(y_1, y_2, ..., y_n)$ occurring.
- For a particular observed sample $(y_1, y_2, ..., y_n)$, we call $f(y_1, y_2, ..., y_n|\theta)$ the **likelihood** of this sample occurring if the true value of the parameters equalled θ.
- The maximum likelihood estimator (MLE) is the estimator θ^{MLE} that maximises the value of the likelihood function for the observed data.
- MLE estimates may be biased but it can be shown that they are consistent and asymptotically efficient, i.e. they have the lowest possible asymptotic variance of all consistent estimators.
- In general, MLEs cannot be obtained using analytical methods, so numerical methods are used to estimate the set of coefficeints that maximise the likelihood function.

MLE with Normal Errors

- Suppose a set of observations $(y_1, y_2, ..., y_n)$ were generated by a normal distribution with an unknown mean and standard deviation, μ and σ. Then the MLEs are the values of μ and σ that maximise the joint likelihood obtained by multiplying together the likelihood of each of the observations

$$f(y_1, y_2, ..., y_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

- Econometricians often work with the log of the likelihood function (this is a monotonic function so maximising the log of the likelihood produces the same estimator). In this example, the log-likelihood is

$$\log f(y_1, y_2, ..., y_n|\mu, \sigma) = -\frac{n}{2}\log 2\pi - n\log\sigma + \sum_{i=1}^{n}\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

- More generally, the log-likelihood function of n observations, $y_1, y_2, ..., y_n$, of a vectors of size k drawn from a $N(\mu, \Sigma)$ multivariate normal distribution is

$$\log f(y_1, y_2, ..., y_n|\mu, \Sigma) = -\frac{(kn)}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{k=1}^{n}(y_i - \mu)'\Sigma^{-1}(y_i - \mu)$$

MLE of Time Series with Normal Errors

- Suppose $\epsilon_t \sim N\left(0, \sigma^2\right)$. Consider the AR(1) model

$$y_t = \rho y_{t-1} + \epsilon_t$$

- Figuring out the joint unconditional distribution of a series $y_1, y_2, ..., y_n$ is tricky but we can say $y_2 \sim N\left(\rho y_1, \sigma^2\right)$ and $y_3 \sim N\left(\rho y_2, \sigma^2\right)$ and so on.
- Let $\theta = (\rho, \sigma)$. Conditional on the first observation, we can write the joint distribution as

$$\begin{aligned} f\left(y_2, ..., y_n | \theta, y_1\right) &= f\left(y_n | \theta, y_{n-1}\right) f\left(y_{n-1} | \theta, y_{n-2}\right) ... f\left(y_2 | \theta, y_1\right) \\ &= \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\left(y_i - \rho y_{i-1}\right)^2}{2\sigma^2}\right] \end{aligned}$$

- The log-likelihood is

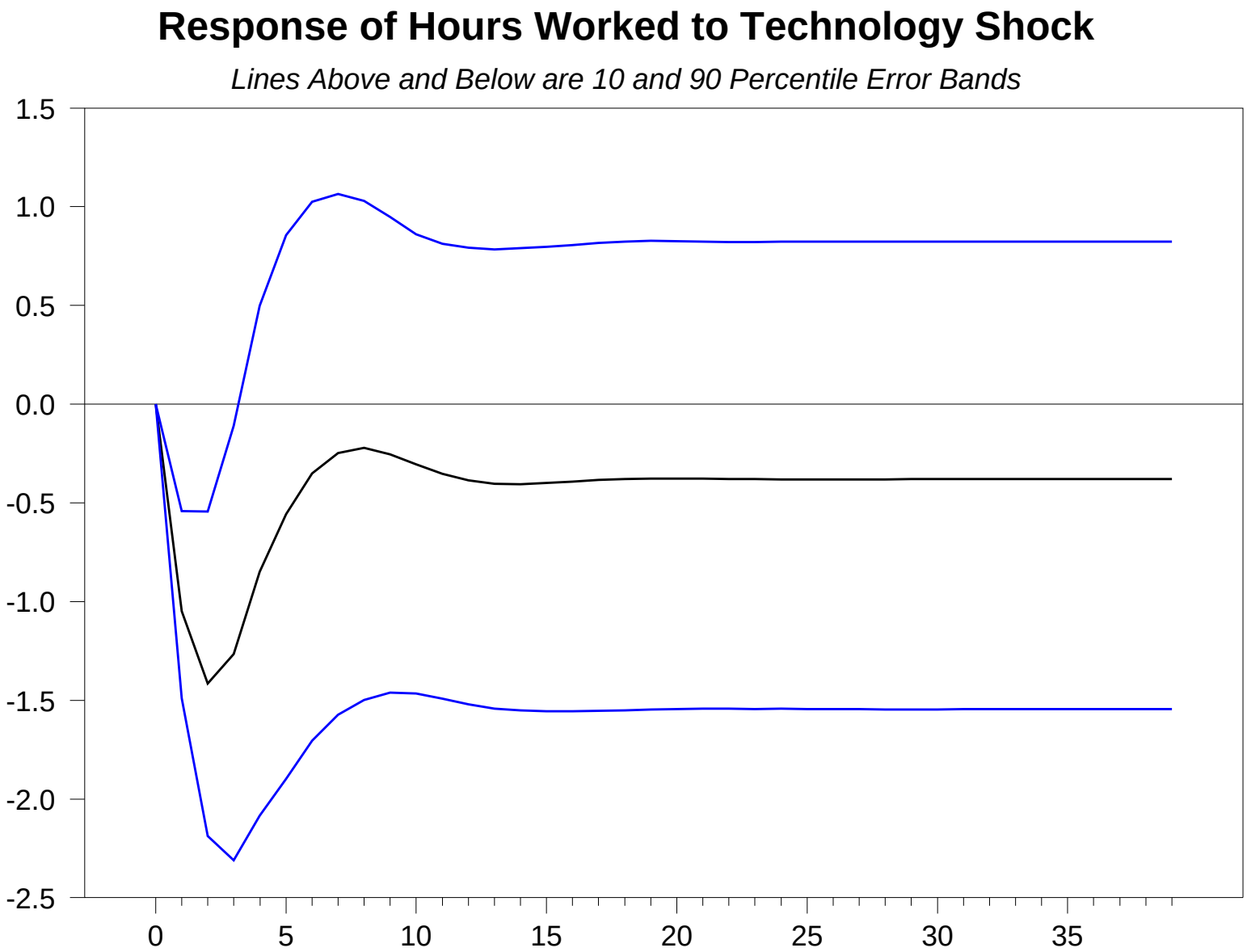
$$\log f\left(y_1, y_2, ..., y_n | \theta, y_1\right) = -\left(n-1\right)\left(\frac{\log 2\pi}{2} + \log\sigma\right) + \sum_{i=2}^{n}\left[\frac{-\left(y_i - \rho y_{i-1}\right)^2}{2\sigma^2}\right]$$

- Can easily show OLS provides the MLE for ρ. Generalises to VAR estimation.

Standard Error Bands for Impulse Response Functions

- When you estimate a VAR model such as $Z_t = AZ_{t-1} + \epsilon_t$ via OLS to obtain estimates $\hat{A}$, you can use the estimated coefficients in $\hat{A}$ to construct estimates of impulse responses $I, \hat{A}, \hat{A}^2, \hat{A}^3, ...$
- As with the underlying coefficient estimates, the estimated IRFs are just our best guesses. We would like to know how confident we can be about them. For example, how sure can we be that the response of a variable to a shock is always positive?
- For this reason, researchers often calculate confidence intervals for each point in an impulse response graph.
- Graphs in VAR papers thus usually show three lines for each impulse response function: The estimated set of impulse responses at each horizon and lines above and below showing the upper and lower ends of some confidence interval. See example on next page.
- For some reason, some researchers sometimes show plus or minus one standard deviation. However, if you want to be confident that responses have a particular sign, then 10th and 90th percentiles (or 5th and 95th) are a better idea.

Example of a Chart With IRF Error Bands



Bootstrapping Standard Errors for IRFs

- Analytical results can be derived to obtain asymptotic (i.e. large sample) distributions for impulse response functions. Unfortunately, these distributions are not very accurate representations of the actual distributions obtained in finite samples of the size used in most empirical work.
- Bootstrap methods are now commonly to derive the standard error bands for IRFs.
- In practice, this is done as follows:
  1. Estimate the VAR via OLS and save the errors $\hat{\epsilon}_t$.
  2. Randomly sample from these errors to create, for example, 10,000 simulated data series $Z_t^* = \hat{A}Z_{t-1}^* + \epsilon_t^*$.
  3. Estimate a VAR model on the simulated data and save the 10,000 different IRFs associated with these estimates.
  4. Calculate quantiles of the simulated IRFs, e.g. of the 10,000 estimates of the effect in period 2 on variable i of shock j.
  5. Use the n-th and $(100 - n)$-th quantiles of the simulated IRFs as confidence intervals.

A Problem: Lots of Parameters

- One problem with classical estimation of VAR systems is that there are lots of parameters to estimate.
- Estimating a Cholesky decomposition VAR with n variables with k lags involves direct estimation of $n^2k + \frac{n(n-1)}{2}$ parameters.
- For 3 variable VAR with one lag, this is already 12 parameters.
- Consider a 6 variable VAR with 6 lags: $(36)(6) + (6)(5)/2 = 231$ coefficients.
- Because many of the coefficients are probably really zero or close to it, this can lead to a severe "over-fitting" problem that can result in poor-quality estimates and bad forecasts.
- This problem can lead researchers to limit the number of variables or number of lags used, perhaps resulting in mis-specification (leaving out important variables or missing important dynamics.)
- This can also lead to poor inference and bad forecasting performance.
- The **Bayesian** approach to VARs deals with this problem by incorporating additional information about coefficients to produce models that are not as highly sensitive to the features of the particular data sets we are using.

Bayes's Law

- Bayes's Law is a well-known result from probability theory. It states that

$$\Pr(A \mid B) \propto \Pr(B \mid A)\, Pr(A)$$

- For example, suppose you have prior knowledge that A is a very unlikely event (e.g. an alien invasion). Then even if you observe something, call it B, that is likely to occur if A is true (e.g. a radio broadcast of an alien invasion), you should probably still place a pretty low weight on A being true.
- In the context of econometric estimation, we can think of this as relating to variables Z and parameters θ. When we write

$$\Pr(\theta = \theta^* \mid Z = D) \propto \Pr(Z = D \mid \theta = \theta^*)\, Pr(\theta = \theta^*)$$

we are calculating the probability that the a vector of parameters θ takes on a particular value, θ^* given the oberved data, D, as a function of two other probabilities: (i) the probability that $Z = D$ if it was the case that $\theta = \theta^*$ and (ii) the probability that $\theta = \theta^*$.

Bayesian Probability Density Functions

- Since coefficients and data in VARs are continuous, we need to write the Bayes relationship in form of probability density functions:

$$f_\theta\left(\theta^* \mid D\right) \propto f_Z\left(D \mid \theta^*\right) f_\theta\left(\theta^*\right)$$

- The function $f_Z\left(D \mid \theta^*\right)$ is the likelihood function—for each possible value of θ^*, it tells you the probability of a given dataset occurring if the true coefficients $\theta = \theta^*$.
- The likelihood functions can be calculated once you have made assumptions about the distributional form of the error process.
- Bayesian analysis specifies a "'prior distribution", $f_\theta(\theta^*)$, which summarises the researcher's pre-existing knowledge about the parameters θ.
- This is combined with the likelihood function to produce a "posterior distribution" $f_\theta\left(\theta^* \mid D\right)$ that specifies the probability of all possible coefficient values given both the observed data and the priors.
- Posterior distributions cannot generally be calculated analytically. Recent progress in computing power and numerical algorithms (via "Markov Chain Monte Carlo" algorithms — see the paper on the website for a discussion) have made Bayesian methods easier to implement.

Bayesian Estimation

- An obvious way to derive a "best estimator" from the posterior distribution is to calculate the mean of the distribution:

$$\hat{\theta} = \int_{-\infty}^{\infty} x f_{\theta}\left(x \mid D\right) dx$$

- You can show that this estimator is a weighted average of the "maximum likelihood estimator" and the mean of the prior distribution, where the weights depend on the covariances of the likelihood and prior functions: The more confidence the researcher specifies in the prior, the more weight will be placed on the prior mean in the estimator.
- With normally distributed errors, the maximum likelihood estimates are simply the OLS estimates, so Bayesian estimators of VAR coefficients are weighted averages of OLS coefficients and the mean of the prior distribution.

Bayesian VARs

- Typically, researchers specify priors so that coefficients are expected to get smaller for longer-lagged variables and that cross-equation coefficients (e.g. effect of lagged X_2 on X_1) are smaller than own-lag effects.
- A common approach is to set the mean of the prior probability distribution for the first own-lag coefficient to be a large positive figure while setting the prior mean for all other coefficients to be zero (e.g. the "Minnesota prior").
- The researcher must also decide how confident they are in this prior e.g. how quickly the prior probabilities move towards zero as you move away from the prior mean. The "tighter" the prior, the higher will be the weight on the prior in calculating the posterior Bayesian estimator.
- This sounds sort of complicated but in practice these days it is not. Various computer packages make it easy to specify priors of a particular form, with the tightness usually summarised by a couple of parameters.
- Unlike models estimated by OLS, Bayesian models with Minnesota-style priors are likely to have most coefficients be close to zero, so they are more parsimonious and less subject to over-fitting problems. But they achieve this without arbitrarily setting parameters to zero as would be done in non-VAR models.

Large Bayesian VARs

- Because of the problem with having to estimate so many parameters, most VAR papers have tended to use a small number of macroeconomic variables.
- However, economists in engaged in forecasting tend to look at a huge range of variables that are available at a monthly or weekly frequency.
- For instance, someone forecasting the US economy may look at employment, unemployment claims, personal income and consumption, industrial production, durable goods orders, figures on inventories, trade data, incoming fiscal data, sentiment surveys and indicators from financial markets.
- All could be useful for forecasting but a 20 variable monthly VAR with 12 lags could not be estimated by traditional methods. How can such a data set be cut down to a few variables without losing valuable information?
- Banbura, Giannone and Reichlin (2008) show that standard Bayesian VAR methods work very well for forecasting with VAR systems that incorporate large numbers of variables, provided that the tightness of the priors is increased as more variables are added.

Banbura-Giannone-Reichlin Evidence on Forecasting

Table 1: Relative MSFE, BVAR

		SMALL	CEE	MEDIUM	LARGE
h=1	EMPL	1.14	0.67	0.54	0.46
	CPI	0.89	0.52	0.50	0.50
	FFR	1.86	0.89	0.78	0.75
h=3	EMPL	0.95	0.65	0.51	0.38
	CPI	0.66	0.41	0.41	0.40
	FFR	1.77	1.07	0.95	0.94
h=6	EMPL	1.11	0.78	0.66	0.50
	CPI	0.64	0.41	0.40	0.40
	FFR	2.08	1.30	1.30	1.29
h=12	EMPL	1.02	1.21	0.86	0.78
	CPI	0.83	0.57	0.47	0.44
	FFR	2.59	1.71	1.48	1.93
λ		∞	0.262	0.108	0.035

Notes: Table reports MSFE relative to that from the benchmark model (random walk with drift) for employment (EMPL), CPI and federal funds rate (FFR) for different forecast horizons h and different models. SMALL, CEE, MEDIUM and LARGE refer to the VARs with 3, 7, 20 and 131 variables, respectively. λ is the shrinkage hyperparameter and is set so that the average in-sample fit for the three variable of interest is the same as in the SMALL model estimated by OLS. The evaluation period is 1971-2003.