Bayesian VAR models

Kevin Kotzé

Table of contents

Introduction

- There are many reasons to make use of Bayesian methods
- Useful in both economic & financial applications
- Particularly adept at dealing with identification issues, different data sources, misspecification, parameter uncertainty and a number of computational matters
- Classical VARs have a problem with the loss of dof and unit roots
  - Bayesian techniques can provide parameter estimates for models with many variables and relatively little data
  - These methods do not provide biased coefficient estimates when the variables contain unit roots
- BVARs are powerful forecasting tools where more weight may be given to the information of own variable lags

Introduction

Identification Issues

- Multivariate macroeconomic models usually have a large number of parameters
  - Note that $dof = (K^2 \times p) + C$
  - Therefore VAR(4) with 5 variables and a constant has 105 parameters, which requires 27 years of quarterly data
  - Without prior information, it is hard to obtain precise estimates of all the parameters
  - Also useful in State-Space models where all the unobserved variables and parameters are all random variables
- Most macroeconomic datasets have a fairly limited number of observations
- Yet we'd often include a number of variables in the model
- The use of time varying parameters (and similar innovations) eats up additional degrees of freedom
- By imposing restrictions on parameters or shrinking them towards zero, we can deal with the over-parametrisation

Introduction

Other Advantages

- Most macroeconomic variables have unit roots, which provide biased coefficients when using classical inference
  - requires the removal of stochastic trend (when not cointegrated) and the loss of useful information
- BVARs are able to deal with misspecification & uncertainty in a theoretically consistent manner since the parameters are random variables
- Could include different sources of data from other studies in the prior
- Computational methods for this purpose within the Bayesian paradigm are extremely efficient and advanced
- Provides synthesis between calibration and estimation, which is useful in macroeconometric models

The Bayesian Paradigm

- To appreciate the fundamental difference between Bayesian and classical inference:
  - Classical inference assumes a parameter vector, Θ, governs the DGP
  - Hence the objective is to find Θ from the data sample
  - The parameter vector Θ would contain point estimates
  - Probability statements refer to the properties of Θ that would be found in a repeated samples

The Bayesian Paradigm

- Bayesian inference assumes the parameter vector Θ contains random variables
- They are concerned with modelling the researcher's beliefs about Θ, which are expressed by a probability distribution
- Bayesian probability concept is a subjective probability statement that does not require a repeated sampling exercise
- Researcher's beliefs that are derived before they inspect the data are summarized by a prior probability distribution
- Information contained in the data is then summarised by the likelihood function of the model
- Together the prior and the likelihood function form the posterior probability distribution
- Posterior conveys everything the researcher knows about the model parameters after having looked at the data

Prior, Likelihood, Posterior

- Bayesian treats the data, $y_t = \{y_1, \ldots, y_T\}$, as given
- Parameters of interest are unknown and inference is conditional on the data
- Prior information on Θ is assumed to be available in the form of a density, $g(\Theta)$
- Density for the DGP conditional on a particular value of the parameter Θ is $f(y|\Theta)$
- This is algebraically identical to the likelihood function $\ell(\Theta|y)$
- Combining these densities where we apply the Bayes rule states that the joint density is

$$f(\Theta, y) = g(\Theta|y)f(y),$$

- Could also derive a similar expression for $f(\Theta, y) = f(y|\Theta)g(\Theta)$

Prior, Likelihood, Posterior

- After equating the conditions for the joint density, we arrive at the conditional probabilities:

$$g(\Theta|y) = \frac{f(y|\Theta)g(\Theta)}{f(y)}$$

- Suggests we are seeking an unknown, Θ parameters, given something that is obtainable, which is the data in y
- We can ignore the term $f(y)$ since it does not involve Θ (which is of interest)
- This enables us to use the expression,

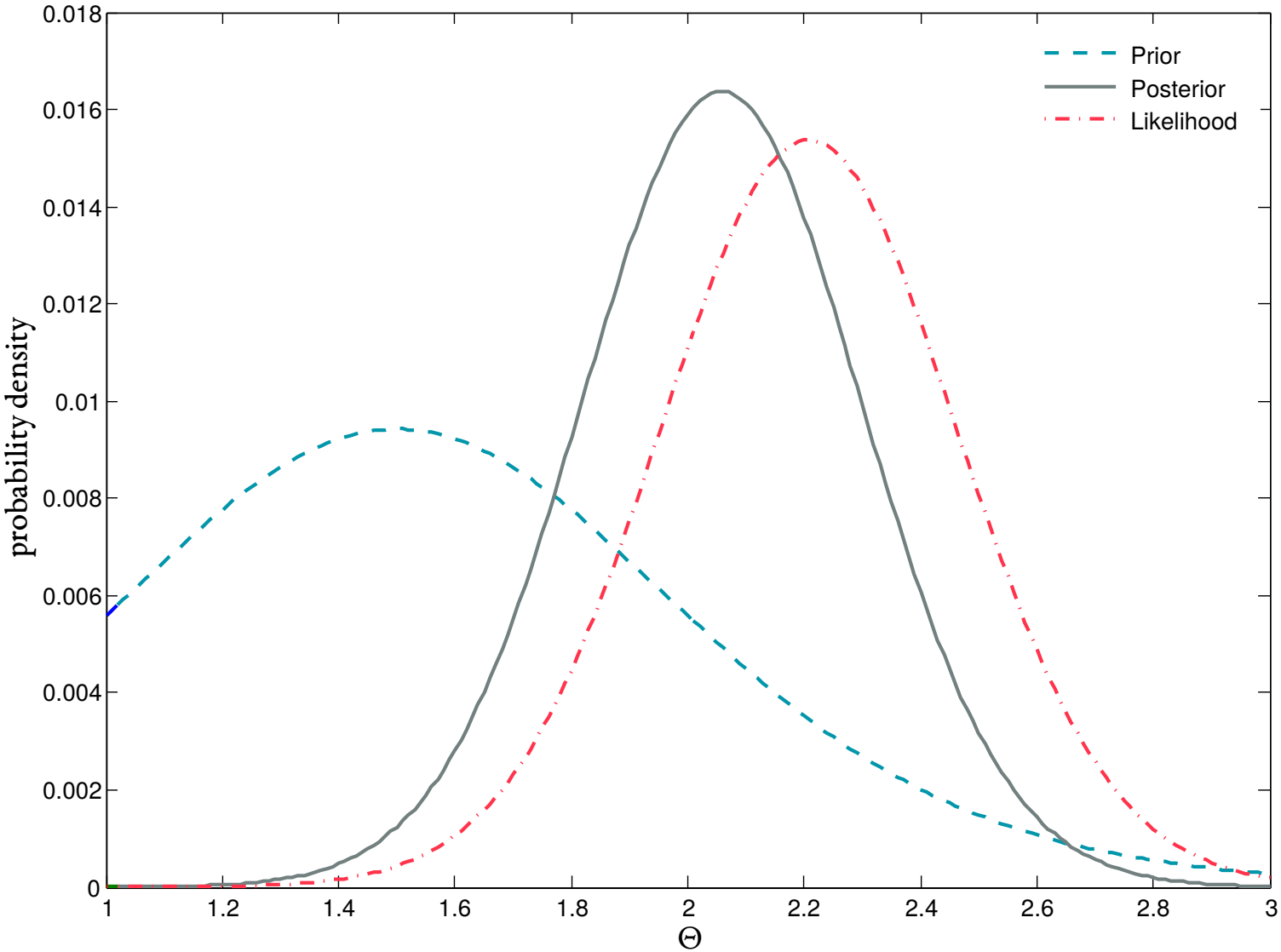
$$g(\Theta|y) \propto f(y|\Theta)g(\Theta) = \ell(y|\Theta)g(\Theta)$$

- where the term $\propto$ is the sign for *'proportional to'*

Prior, Likelihood, Posterior

- Note that:
- $g(\Theta|y)$ is the **posterior density** - probability attached to particular parameter values after observing the data
- $\ell(y|\Theta)$ is the **likelihood function** - density of the data conditional on the parameters in the model
- $g(\Theta)$ is the **prior density** - what we know about Θ prior to seeing the data
- Hence, to obtain estimates for Θ given the data (i.e. $g(\Theta|y)$), we need to multiply our prior knowledge by a likelihood function that is used to describe the data generating process
- This computation may be extremely intensive as it usually involves taking the product of distributions
- Thereafter the joint distribution is sampled by a subsequent algorithm to derive the individual parameter estimates

Prior, Likelihood, Posterior



Prior, Likelihood, Posterior

- Above figure contains an example of a prior with a mean of 1.5 and a relatively large standard deviation to provide a reasonably flat density
- The likelihood function is associated with higher degrees of certainty and is centred around a mean value of 2.3
- Note that the posterior has a mean 2.1, which needs to be closer to the likelihood function
- The density of the posterior provides us with information about the probability of the coefficient taking on values between 1 and 3
- This would allow us to work out the exact probability of any value within this range rather than the probability of observing a single point estimate

Priors for Reduced-Form VAR Parameters

- It is convenient to specify the prior such that the posterior is from a known family of distributions
- If the prior and the likelihood function have the same functional form then we have a conjugate prior and the posterior will have a similar distribution
- For example, if the likelihood is Gaussian and the prior is also normal, then the posterior again has a normal distribution
- When using priors from a known family of distributions we can impose additional structure on the prior to reduce the number of parameters to a few hyperparameters

Minnesota Prior

- Litterman (1986) and Doan, Litterman, and Sims (1984) propose a specific Gaussian prior for the parameters
- The original proposal shrinks the VAR estimates toward a multivariate random walk model
- Where it is assumed that the underlying data is $I(1)$
- This practice has been found to be useful when forecasting persistent economic time series variables
- In each equation, set the prior mean of the first lag of the dependent variable to one and set the prior mean of all other coefficients to zero
- In other words, if the prior means were the true parameter values, each variable would follow a random walk

Minnesota Prior

- Thereafter, set the prior variances of the intercept terms to infinity as we have very little information about these
- The prior variance of the ij^{th} elements of A_p, denoted as ij,p, to

$$\upsilon_{ij,p} = \begin{cases} (\lambda/p)^2 & \text{if } i = j, \\ (\lambda\theta\sigma_i/p\sigma_j)^2 & \text{if } i \neq j, \end{cases}$$

- where λ is the prior standard deviation of $a_{ii,1}$
- $0 < \theta < 1$ controls the relative tightness of the prior variance in the other lags
- σ_i^2 is the i^{th} diagonal element of Σ_u

Minnesota Prior

- For example, in a bivariate VAR(2) model with all the coefficients evaluated at their prior mean

$$
\begin{aligned}
y_{1,t} = \underset{(\infty)}{0} + \underset{(\lambda)}{1 \cdot y_{1,t-1}} + \underset{(\lambda\theta\sigma_1/\sigma_2)}{0 \cdot y_{2,t-1}} + \underset{(\lambda/2)}{0 \cdot y_{1,t-2}} + \cdots \\
+ \underset{(\lambda\theta\sigma_1/2\sigma_2)}{0 \cdot y_{2,t-2}} + u_{1,t} \, , \\
y_{2,t} = \underset{(\infty)}{0} + \underset{(\lambda\theta\sigma_2/\sigma_1)}{0 \cdot y_{1,t-1}} + \underset{(\lambda)}{1 \cdot y_{2,t-1}} + \underset{(\lambda\theta\sigma_2/2\sigma_1)}{0 \cdot y_{1,t-2}} + \cdots \\
+ \underset{(\lambda/2)}{0 \cdot y_{2,t-2}} + u_{2,t} \, .
\end{aligned}
$$

- Here the numbers in parentheses are the prior standard deviations

Minnesota Prior

- Each of the two equations specifies a random walk prior mean for the dependent variables
- The nonzero prior standard deviations reflect the uncertainty regarding the validity of that model
- The standard deviations decline with increasing lag length because more recent lags are assumed to be more likely to have nonzero values
- The advantage of the Minnesota prior is that it reduces the problem of specifying a prior to one of selecting two parameters
- Researcher has to choose only the two hyperparameters λ and θ
- λ controls the overall prior variance of all VAR coefficients
- Smaller λ imply a stronger shrinkage towards the prior mean
- θ controls the tightness of the variances of the coefficients of lagged variables other than the dependent variable
- Values for θ close to one imply that all coefficients of lag 1 have about the same prior variance
- For example, Litterman (1986) finds that $\theta = 0.3$ and $\lambda = 0.2$ works well when forecasting U.S. macroeconomic variables

Minnesota Prior

Motivation for Shrinkage

- If the above model represents a classical VAR:
- And the coefficient $\alpha_{12,2}$ has a large coefficient value and a large standard error
- Despite the large standard error, the large coefficient will still have a significant effect on the IRFs, forecasts, etc.
- In the Bayesian VAR, the effect of such a coefficient would be small, provided that the prior mean for this particular coefficient is close to zero
    - In this case the likelihood would be relatively flat (uninformative) and the posterior would closely approximate the prior
- We would say that this coefficient value shrinks to zero in the BVAR

Minnesota Prior

- Other practical problems that have to be addressed include the specification for the elements of Σ_u
- A simple alternative is to replace Σ_u by its LS estimator or its ML estimator
- This is not particularly desirable as Σ_u is not treated as a purely unknown parameter, thereby ignoring the measure of uncertainty that would have been generated during the estimation procedure
- As an alternative researchers have made use of natural conjugate Gaussian-Inverse Wishart priors or Independent Gaussian-Inverse Wishart priors
- Another potential disadvantage is that it ignores the effects of variables that may be related through a cointegrating vector
- Note that if variables do not have stochastic trends then we may wish to shrink all the coefficients to zero
- Many other modifications of the Minnesota prior have been proposed

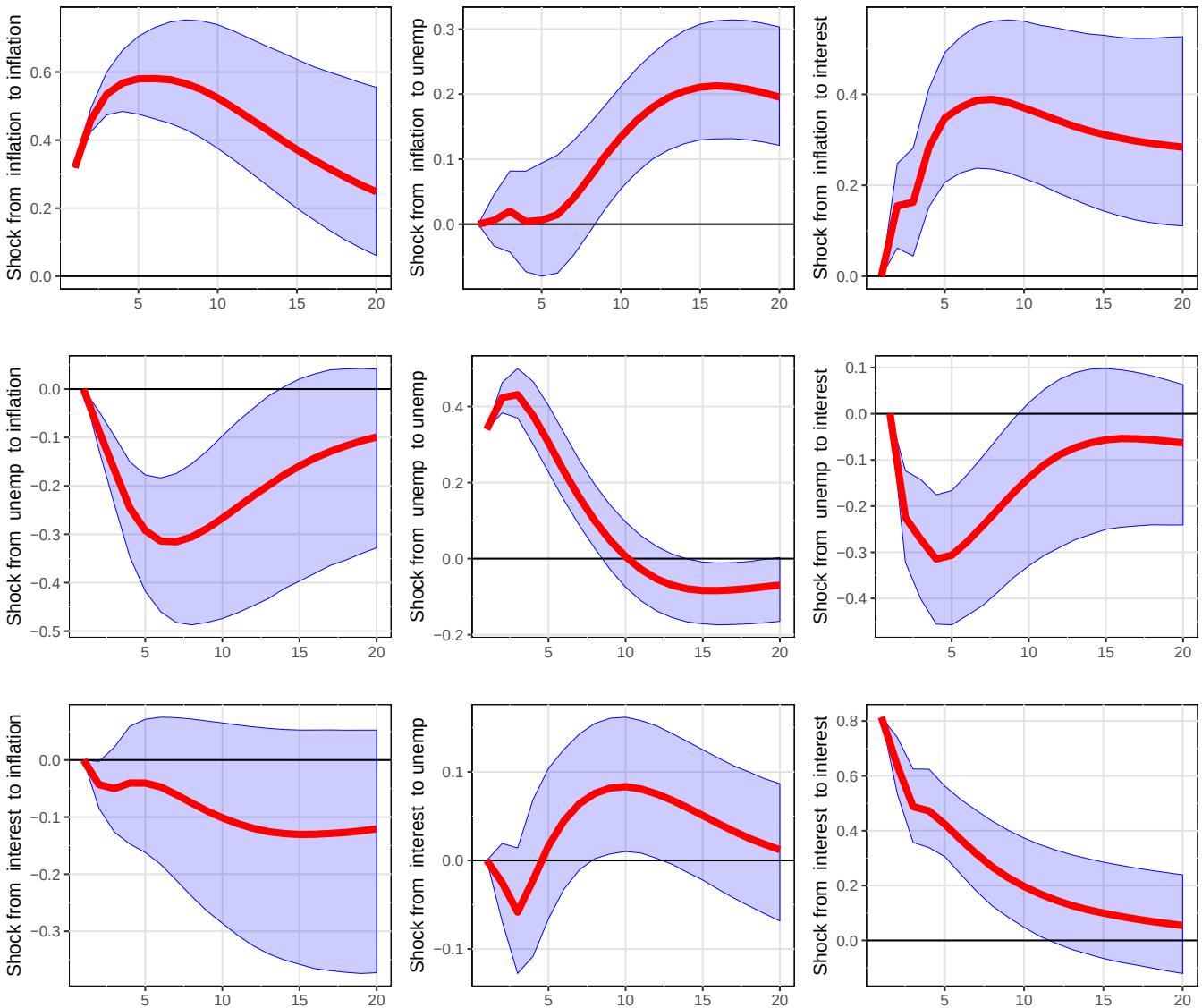
Minnesota Prior

Illustration

- Consider a model with quarterly GDP deflator inflation (π_t), unemployment rate (une_t) and interest rates (r_t)
- Make use of a VAR(4) model with intercept and impose a conventional Minnesota prior
- Unknown error variances are replaced by estimates obtained from fitting univariate AR(4) models to the individual model variables

Minnesota Prior

Illustration



Minnesota Prior

Illustration

- Impact of various shocks on the posterior density are presented in the structural impulse responses
- Apply a recursive structure on the impact multiplier matrix where the variables are ordered $y_t = (\pi_t, une_t, r_t)'$
- Note that the interest rate is ordered last, so the shock to the interest rate equation may be interpreted as a monetary policy shock with no contemporaneous effect on inflation and unemployment
- Suggests that a contractionary monetary policy shock results in a decline in inflation (although the confidence intervals are relatively large)

Models with sign restrictions

- Standard identification of structural shocks usually involves the application of short-run or long-run restrictions as in Sims (1980) and Blanchard & Quah (1989)
- Such restrictions may be inconsistent with several theoretical structures
- Faust (1998), Canova & Nicolo (2002), and Uhlig (2005) present an alternative where they use prior belief about the signs of the impact of certain shocks to identify the structural shocks
- Most common current implementations include the Uhlig (2005) rejection method, Uhlig (2005) penalty function method, Rubio-Ramìrez, Waggoner, & Zha (2010) rejection method, and Fry and Pagan (2011) median target method

Models with sign restrictions

- Consider the following reduced-form VAR(1) model with k endogenous variables

$$y_t = Ay_{t-1} + u_t \text{ for } t = 1, 2, \ldots, T$$

- u_t is a set of errors with mean zero, zero autocorrelation, and variance-covariance matrix

$$\Sigma_u = \mathbb{E}\left[u_t u_t'\right]$$

- u_t is serially uncorrelated and orthogonal to the regressors in each equation and has no structural interpretation since the elements might be correlated across equations

Models with sign restrictions

- It is difficult to impose sign restrictions directly on the coefficient matrix of the model
- However we can impose them ex-post on a set of orthogonalised impulse response functions
- Hence sign restrictions essentially explore the space of orthogonal decompositions of the shocks to see whether the responses conform with the imposed restrictions
- One also needs to make a choice about the length of time these restrictions should apply after the impact of the shock

Models with sign restrictions

- Steps to recover the structural shocks, given a set of sign restrictions, involve:

1. Run an unrestricted VAR in order to get $\widehat{A}$ and $\widehat{\Sigma}_u$, which can be estimated by OLS
2. Extract the orthogonal innovations from the model using a Cholesky decomposition (used to orthogonalise shocks)
3. Calculate the resulting impulse responses
4. Randomly draw an orthogonal impulse vector, which is α
5. Multiply the responses from Step 3 times α and check if they match the imposed signs
6. If yes, keep the response - if not, drop the draw
7. Repeat Steps 2-6

Models with sign restrictions

- The impact multipliers or impulse vector essentially sets the loading of the shock onto the variables
- Uhlig (2005) shows that a vector is an impulse vector, if there is an k-dimensional vector a of unit length, such that:

$$\alpha = \tilde{B}a$$

- where $\tilde{B}\tilde{B}' = \Sigma_u$ is a matrix decomposition of Σ_u
- Uhlig (2005) shows that, given an impulse vector α, one can simply calculate the impulse responses by multiplying the impulse vector with the impulse responses obtained in Step (3)

Models with sign restrictions

- The type of decomposition differ slightly across the three implementations that we consider
- Two procedures suggested by Uhlig (2005) are based on a Givens rotation
- Rubio-Ramìrez, *et al*. (2010) is based on a QR-decomposition
- Moon & Schorfheide (2012) note that sign restrictions are only well defined from a Bayesian point of view
- Steps 2-6 are based on a joint draw from a flat Normal inverted-Wishart posterior for the VAR parameters and a uniform distribution for α
- One can then conduct inference based on the accepted draws and construct error bands similar to Sims & Zha (1999)
- There are a number of reasons for choosing a flat prior for α as discussed in Baumeister & Hamilton (2015) and Uhlig (2005)

Models with sign restrictions

- Uhlig (2005) and Rubio-Ramìrez, *et al*. (2010) rejection methods make use of a number of sub-draws to generate α
- Algorithm then checks whether the imposed sign restrictions are satisfied for each restricted response
- If all restrictions for all restricted periods are met, the draw is kept and the function moves to the next posterior draw - otherwise the draw is rejected
- Uhlig (2005) penalty function seeks to find an impulse vector a that comes as close as possible to satisfying the imposed sign restrictions

Models with sign restrictions

- Original question in Uhlig (2005) was to analyse the effect of an unanticipated monetary policy shock
- Data for the U.S. on Real GDP (y_t), GDP deflator (π_t), commodity price index (p_t), FED funds rate (i_t), non-borrowed reserves (rnb_t), and total reserves (res_t) from 1965q1 to 2003q12
- Suggested that monetary policy shock policy rate:
  - does not decrease the policy rate for x months after the shock
  - does not increase commodity prices for x months after the shock
  - does not increase inflation for x months after the shock
  - does not increase non-borrowed reserves for x months after the shock

Models with sign restrictions

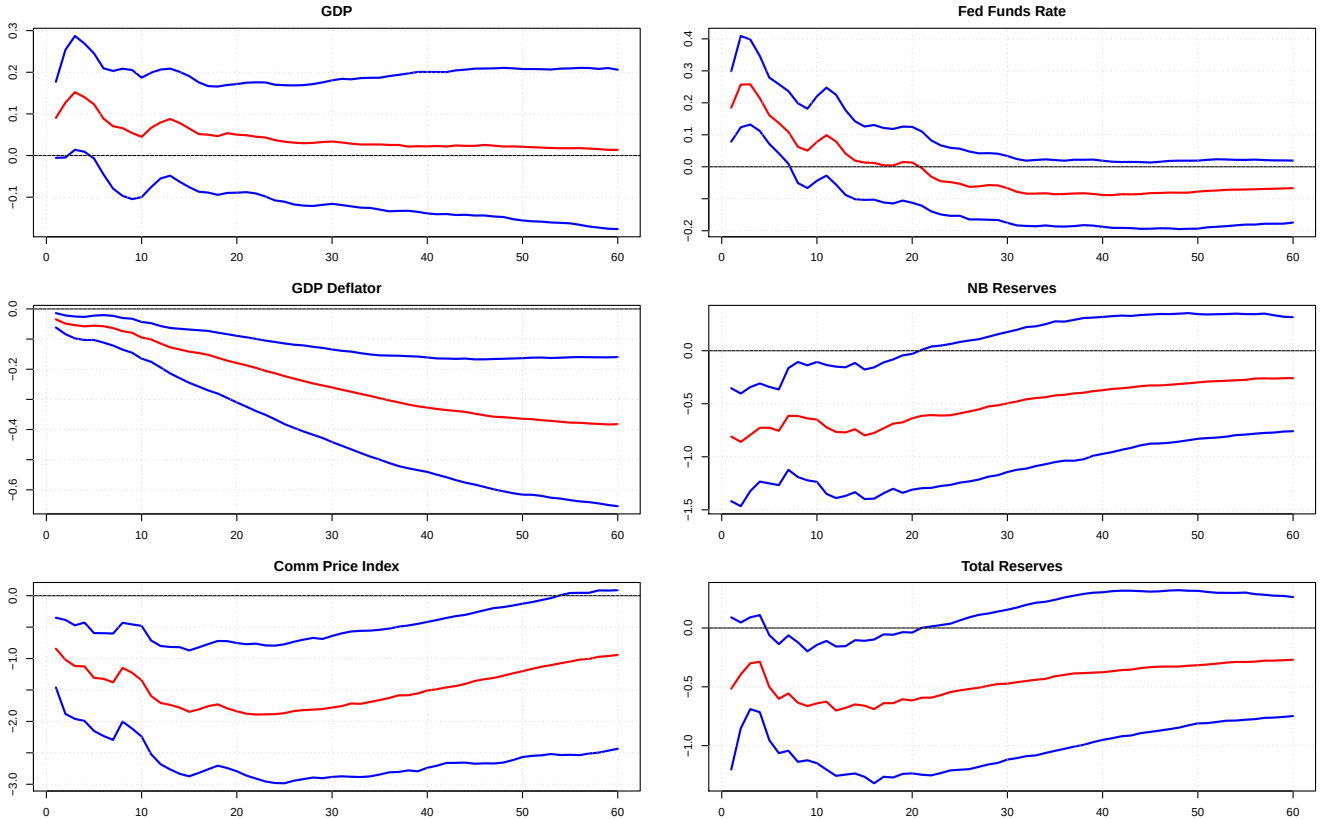
- Given the ordering of the variables, y_t is followed by π_t, p_t, i_t, rnb_t, and res_t, one can specify these restrictions for an interest rate shock

y_t	π_t	p_t	i_t	rnb_t	res_t
$\leqslant 0$	≤ 0	≤ 0	≥ 0	≤ 0	$\leqslant 0$

- Note that the 1st and the 6th variable of the model remain unrestricted (real GDP and total reserves)
- The standard way of analysing the results of a Bayesian VAR model is to take the median of the impulse response draws and plot them together with a pair user specified error bands
- Alternatively one may present the FEVDs of the model

Models with sign restrictions

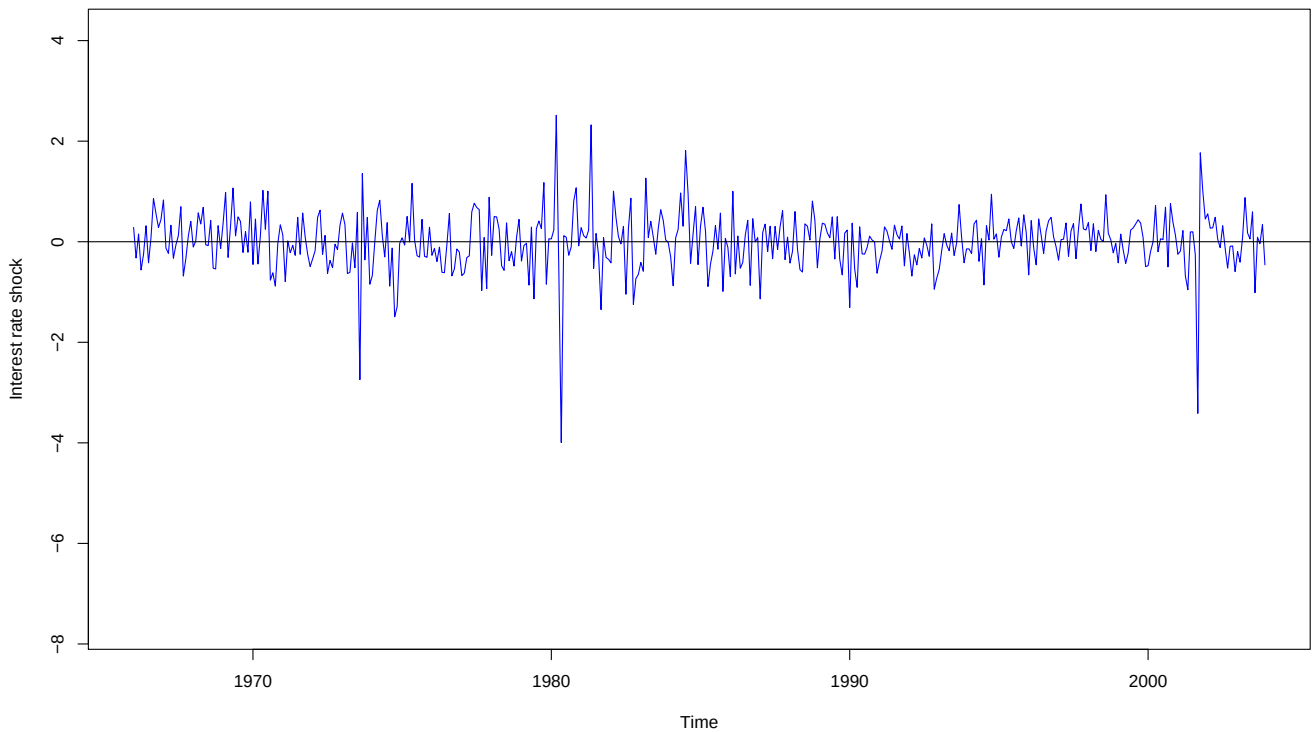
Uhlig (2005) rejection method



- where the direction of the IRF is consistent with theory

Models with sign restrictions

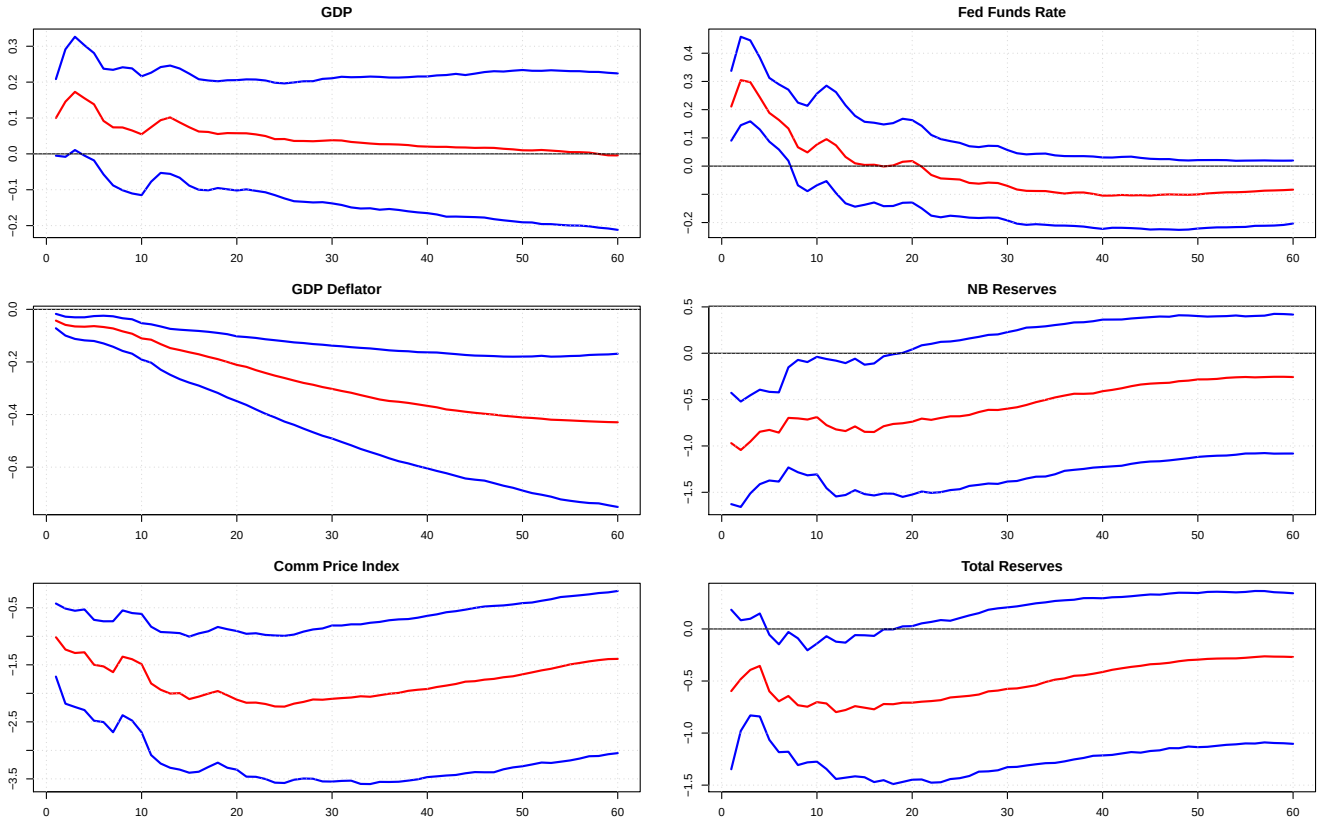
Uhlig (2005) rejection method



- The implied shocks that may be derived from a particular variable

Models with sign restrictions

Rubio-Ramìrez et al. (2010) rejection method



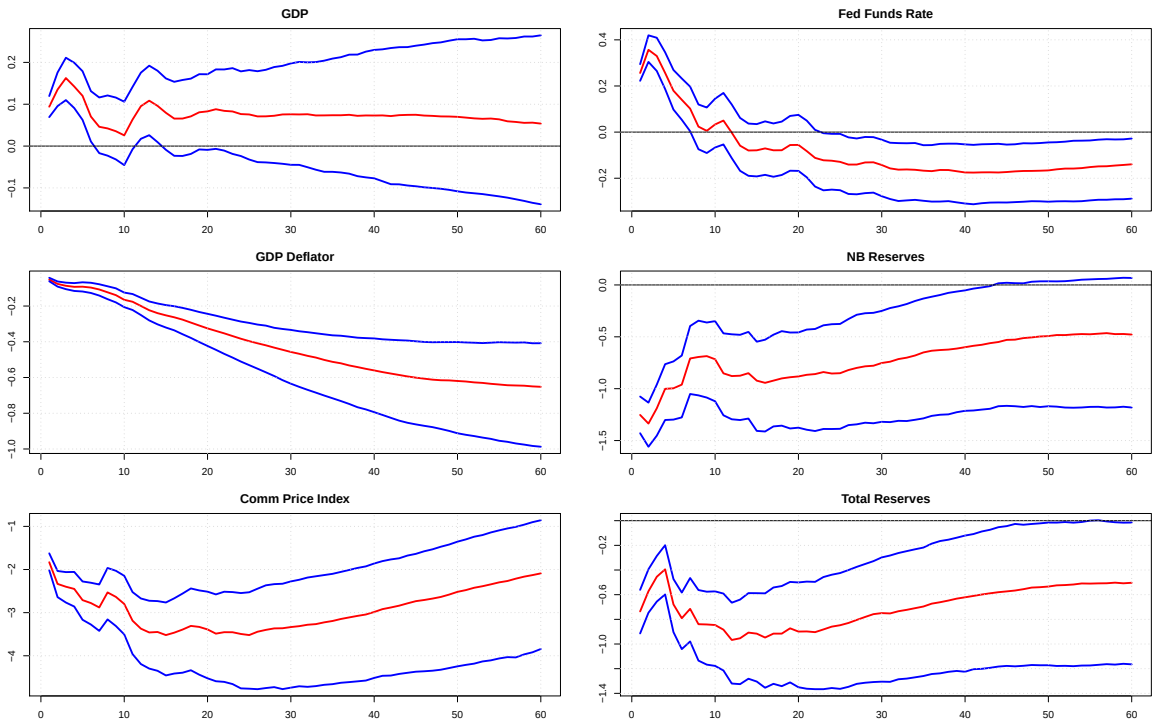
- Rubio-Ramirez et al. (2010) rejection method provides similar results

Models with sign restrictions

- Shortcoming of the two rejection methods is that all impulse vectors satisfying the sign restrictions are considered to be equally likely
- Penalty function method seeks to identify the most likely IRF
- This condition minimises a criterion function that penalises for sign violations

Models with sign restrictions

Uhlig (2005) penalty method



- Again these are highly similar to what we had previously

Models with sign restrictions

- These methods allow for partial identification of the model where we place restrictions on the responses of some variables, but are agnostic about others
- While it is possible to identify all shocks of the model, doing so by just using sign restrictions is inherently difficult
- One reason for this is that different shocks in the model might be characterised by the same set of restrictions
- Thus focussing only on one shock of the model and being explicit about partial identification might be a better way of approaching a particular research question

Models with sign restrictions

- Models identified by sign restrictions are set-identified
- They might not necessarily generate a unique set of impulse responses
- Depending on the problem at hand, sign restrictions may generate a new set of structural equations and shocks for each rotation of α which means each draw produces a set of possible inferences
- The rejection methods are particularly prone to this "model identification" problem

Models with sign restrictions

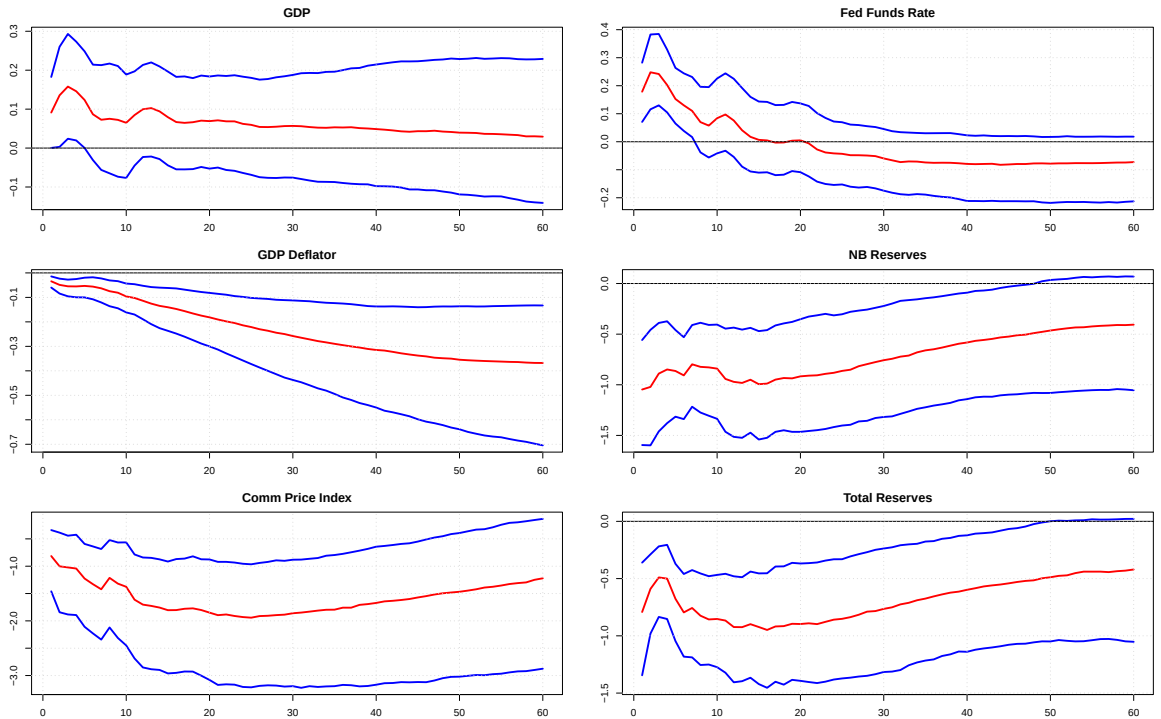
- There are several ways to deal with this problem:
- Fry and Pagan (2011) suggest one should narrow down the set of admissible models to a singleton
- This is what the Uhlig (2005) penalty function does by generating a "weighted sample" of all draws
- Thus, if the results differ markedly across the three routines, the researcher should probably opt for the penalty function specification

Models with sign restrictions

- To improve upon the results of the rejection methods we could use additional information and additional constraints
- In general the more restrictions one imposes the "better" identified the model - provided that the additional restrictions make sense
- Imposing an additional restriction on the response of total reserves to a monetary policy shock yields impulse responses that are closer to the ones produced by the penalty function

Models with sign restrictions

Uhlig (2005) rejection method with additional constraint

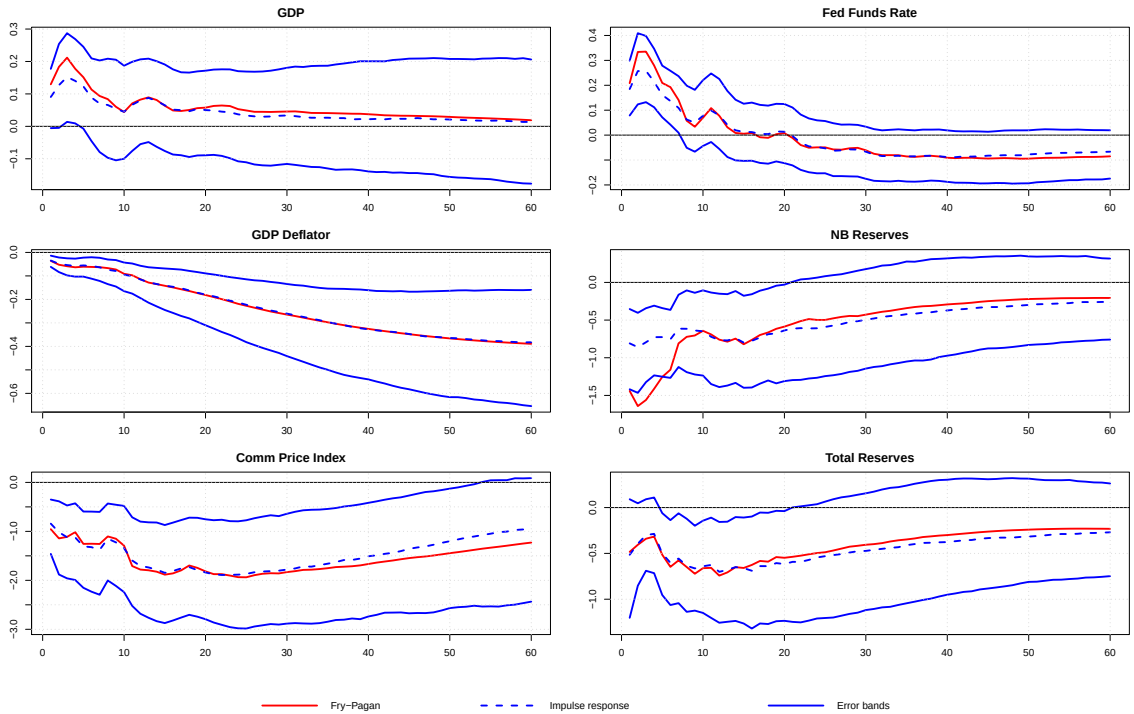


Models with sign restrictions

- Alternatively the Fry and Pagan (2011) Median-Target (MT) method involves finding the single impulse vector that produces impulse responses that are as close to the median responses as possible
- Once could compare the responses of Uhlig (2005) rejection method and the results of Fry & Pagan (2011) MT method in a graph
- Responses are similar in the long run, which may support the evidence in favour of the Uhlig (2005) model specification

Models with sign restrictions

Median-Target method



Conclusion

- Bayesian methods provide consistent way of imposing restrictions on potentially over-parameterised models
- Involves shrinking coefficients towards a particular value (usually zero)
- Also able to deal with the problem of biased coefficients, when in the presence of a unit root
- The use of the Minnesota prior, which provides a convenient methodology for implementing these procedures have generally provided impressive forecasting results

Conclusion

- Thereafter, we considered the estimation of models that make use of sign restrictions
- Models include the rejection and penalty function methods of Uhlig (2005)
- Also considered the results of a model that makes use of the rejection method of Rubio-Ramìrez et al. (2010)
- As well as the Median-Target method of Fry & Pagan (2011)
- When all the results are similar then we can conclude that a single model may be responsible for the correctly identified IRF