Bayesian VAR models

Kevin Kotzé

Contents

1. Introduction
2. The Bayesian Paradigm
3. Priors for Reduced-Form VAR Parameters
4. Models with sign restrictions
5. Conclusion

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

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 would 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

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 \textbf{posterior density} - probability attached to particular parameter values after observing the data
- $\ell(y|\Theta)$ is the \textbf{likelihood function} - density of the data conditional on the parameters in the model
- $g(\Theta)$ is the \textbf{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

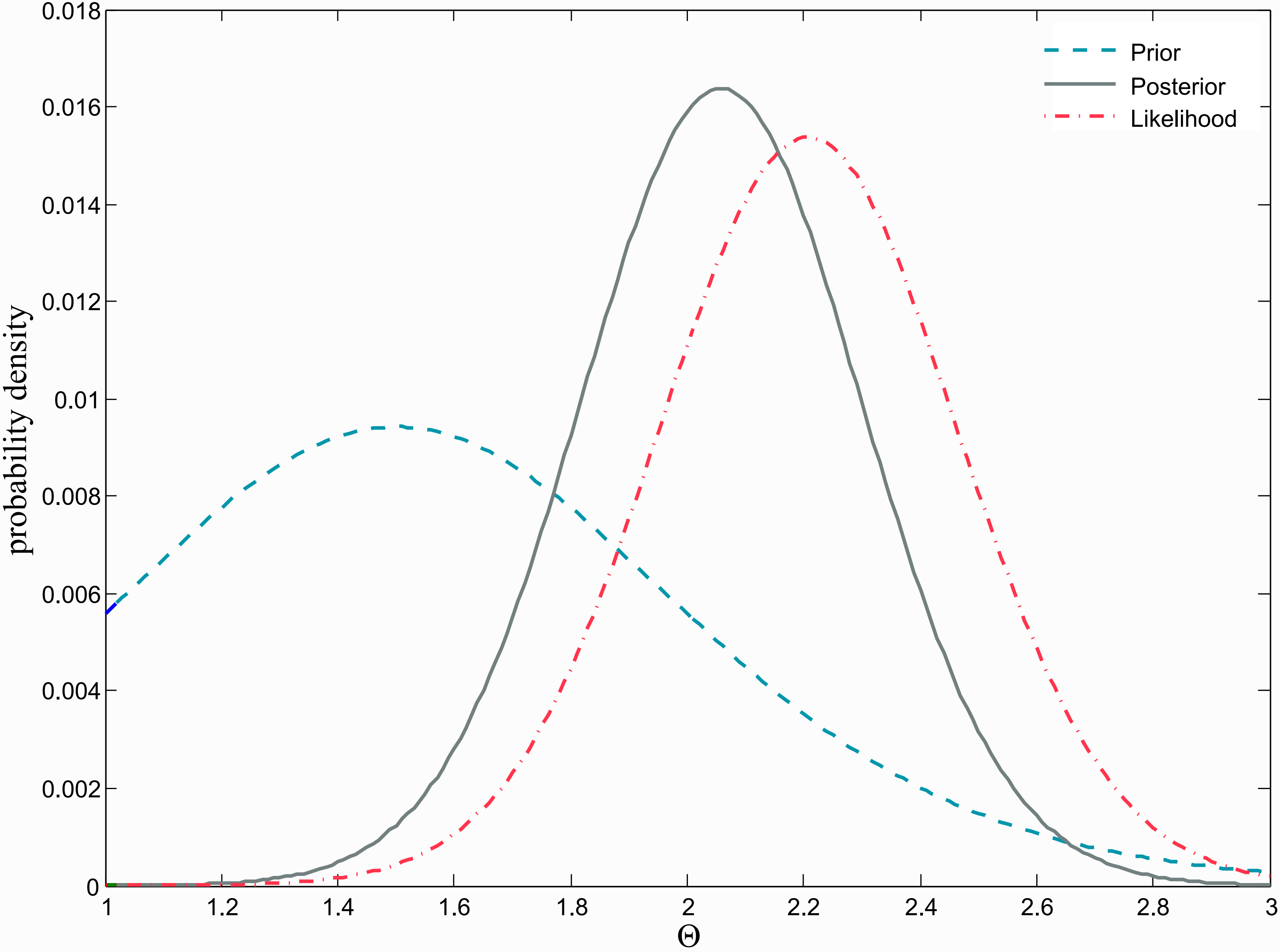

Figure : Prior, Likelihood, Posterior

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

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

- 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{eqnarray} \begin{array}{c} y_{1,t} = \\ \; \end{array} \begin{array}{c} 0\ (\infty) \end{array}\begin{array}{c} +\\ \; \end{array} \begin{array}{c} 1 \cdot y*{1,t-1}*\ *(\lambda) \end{array}*

$$+$$

\begin{array}{c} 0 \cdot y{2,t-1}\ (\lambda \theta \sigma*1 / \sigma_2) \end{array}*

$$+$$

\begin{array}{c} 0 \cdot y{1,t-2}\ (\lambda/2) \end{array}

$$+$$

\ldots \\

$$+$$

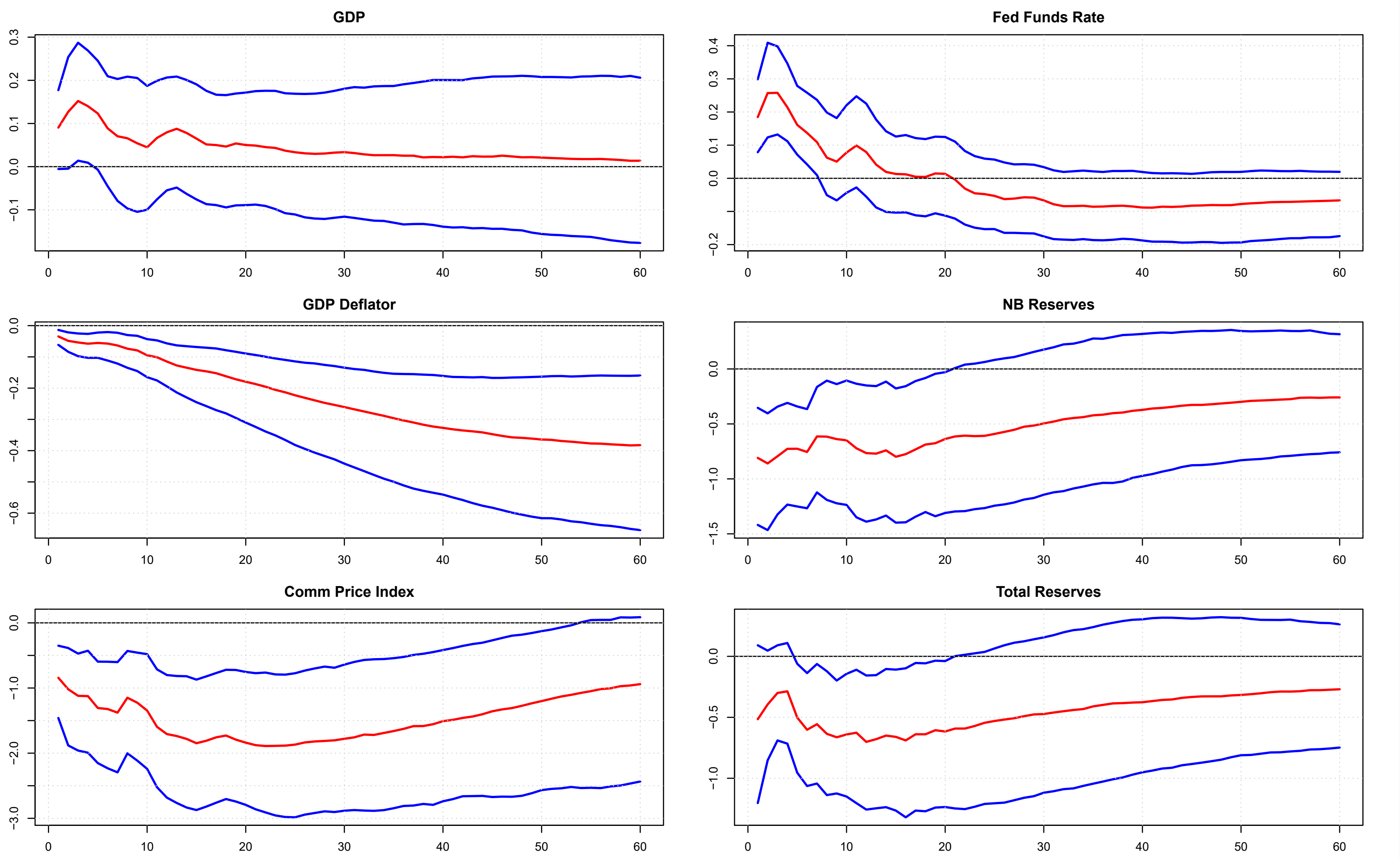

Figure : Uhlig (2005) rejection method

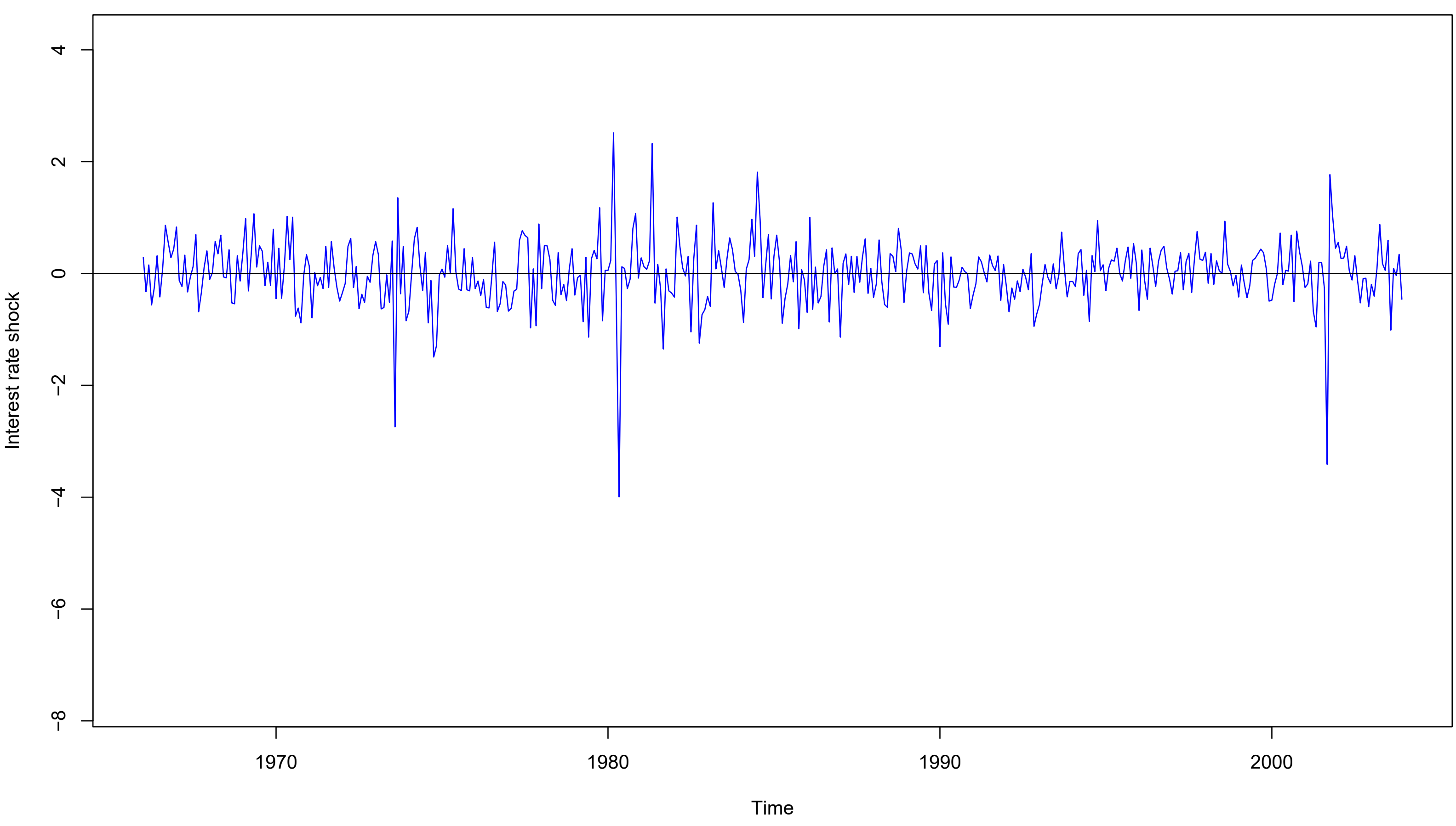

Figure : Uhlig (2005) rejection method

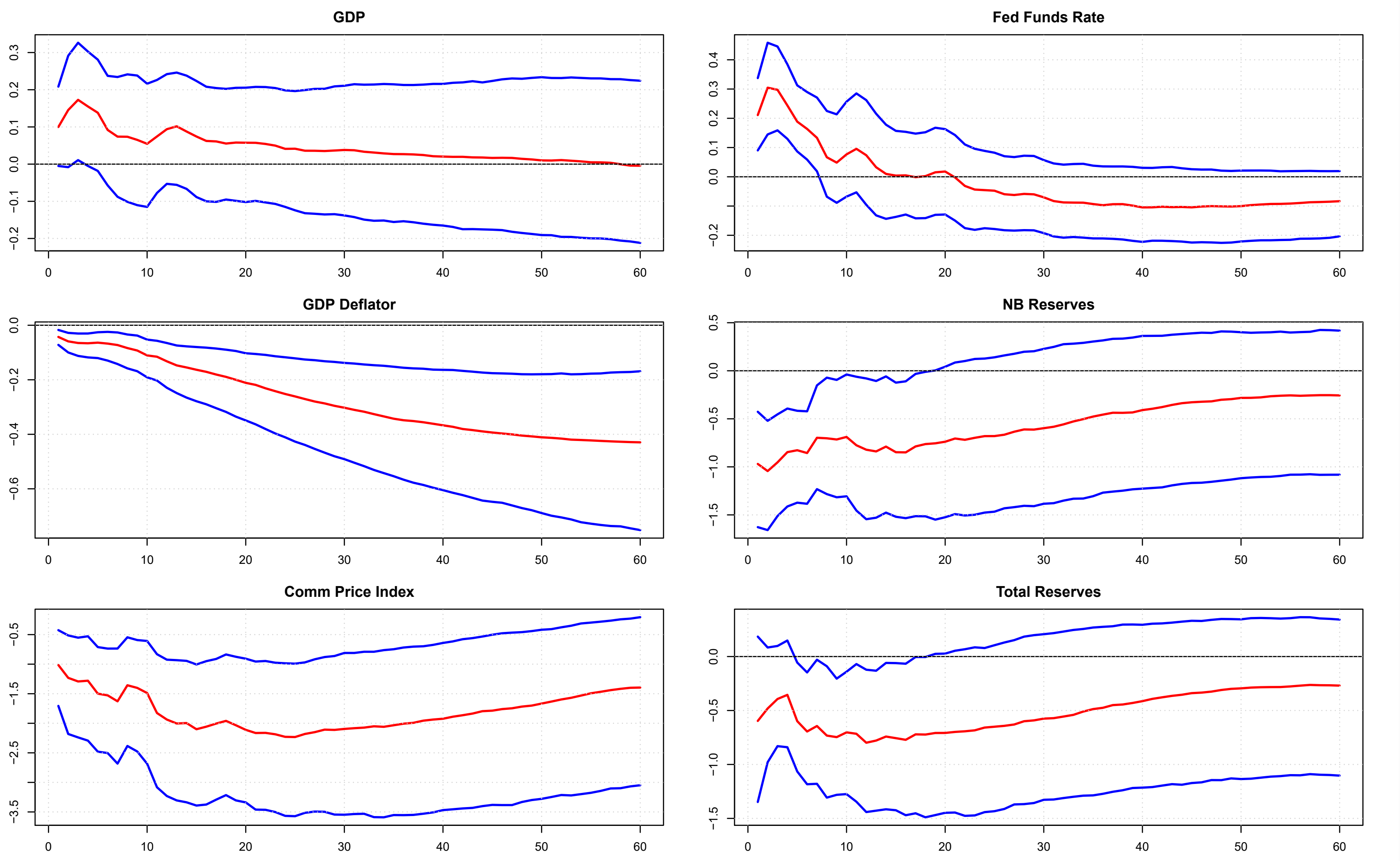

Figure : Rubio-Ramirez et al. (2010) rejection method provides similar results

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

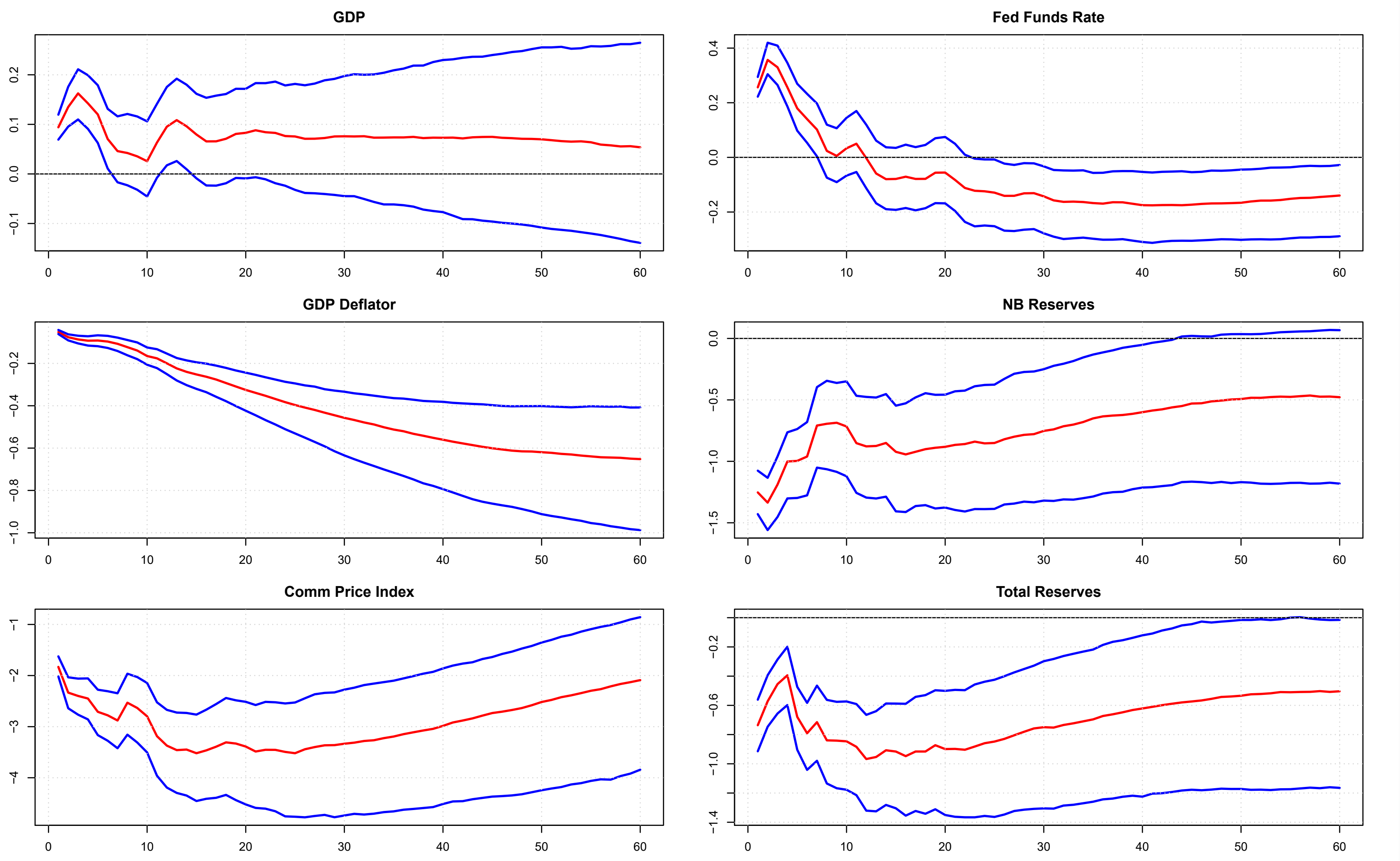

Figure : Uhlig (2005) penalty method

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

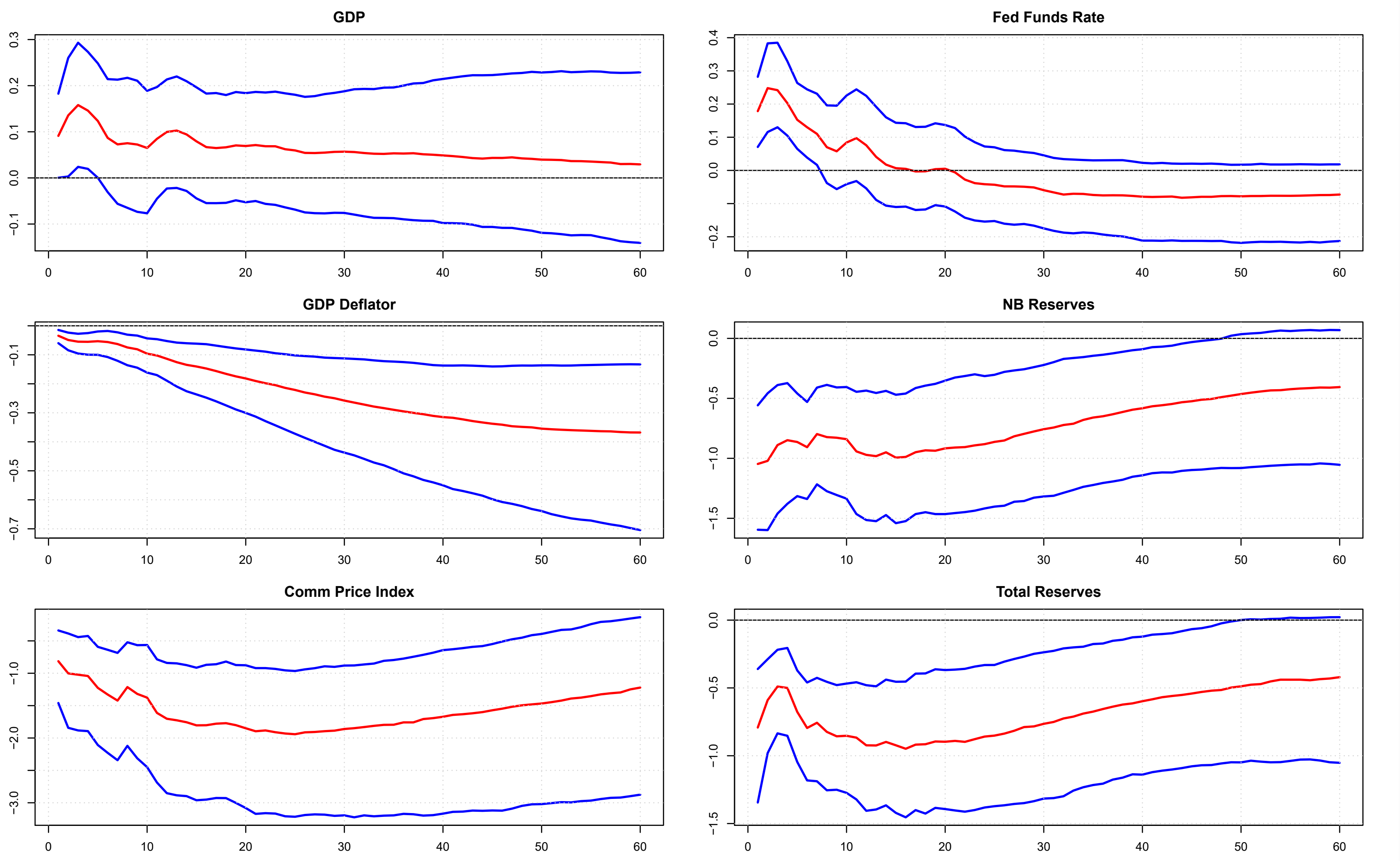

Figure : Uhlig (2005) rejection method with additional constraint

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

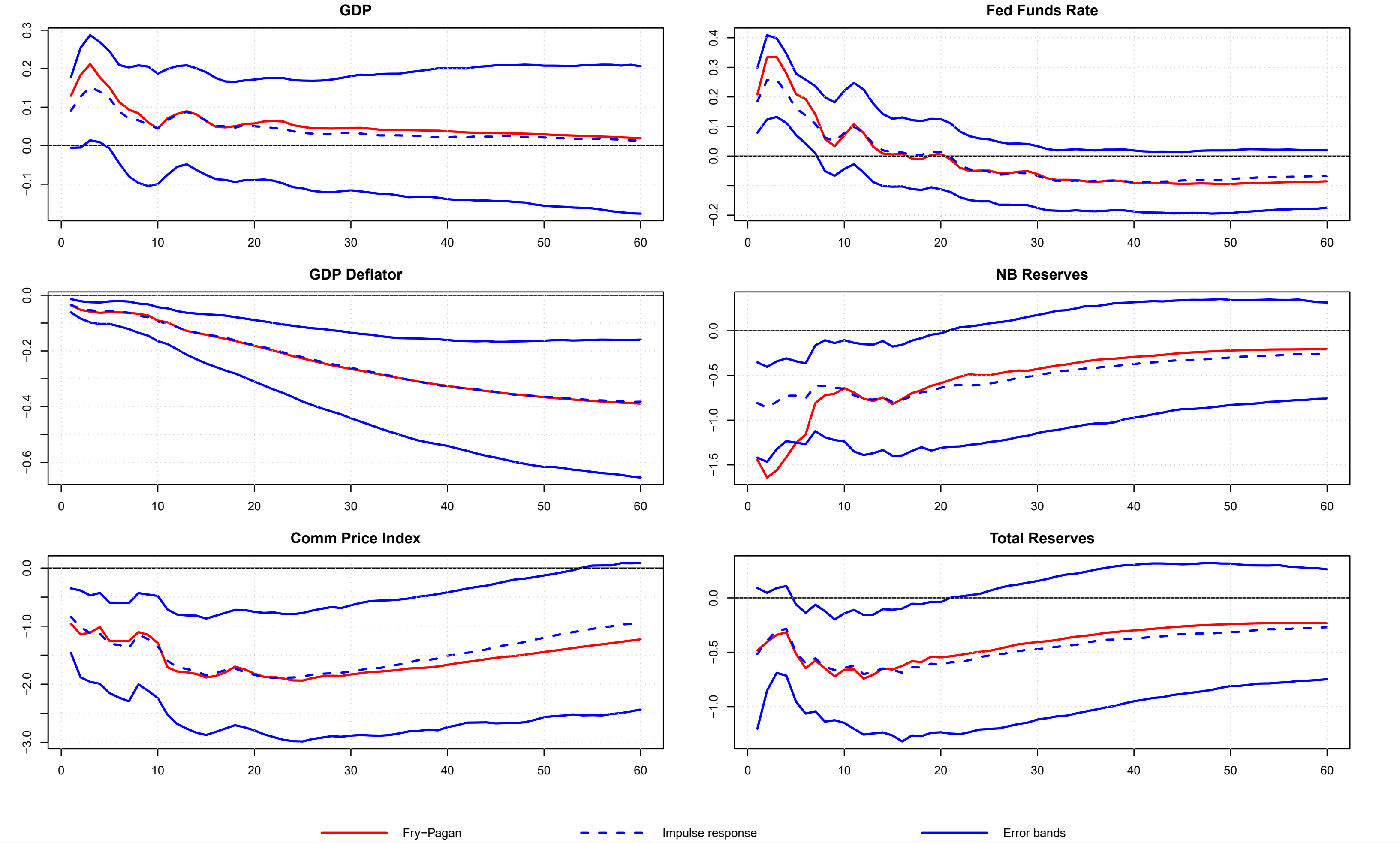

Figure : Median-Target method

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-Ram{\'i}rez 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