Bayesian vector autoregression models

by Kevin Kotzé

During the year of 1763, the Royal Society of London published a posthumous article that described the thoughts of Reverend Thomas Bayes (born circa 1702; died 1761). The paper, which is titled "*An Essay Towards Solving a Problem in the Doctrine of Chances*", established the basis of a method that determines the chance of realising an uncertain event through the accumulation of evidence, using what is now termed *Bayes theorem*.[1] This technique can be used to estimate parameters, in the sense that we can accumulate evidence from data, to determine the chance of realising certain parameter values.

The popularity of the Bayesian approach to econometric modelling continues to enjoy a growing number of followers and many macroeconomists argue that it has many important advantages over the classical (frequentist) approach in a number of instances. Support for this point of view can be found in Schorfheidede and Negro (2011), Fernández-Villaverde, Guerrón-Quintana, and Rubio-Ramìrez (2010), and Koop and Korobilis (2010), among others; where it is suggested that these methods are potentially more adept at dealing with identification issues, different data sources, misspecification, parameter uncertainty and a number of computational matters. These methods are also able to incorporate information from different sources or other studies in the analysis, through the use of a carefully specified prior.

In addition, while classical vector autoregressive (VAR) models have a problem with the loss of degrees Bayesian techniques can be used to provide parameter estimates where the models include many variables and relatively little data. As has already been noted, these methods may also be used to provide coefficient estimates that are not biased, when the variables contain unit roots. Furthermore, when using these techniques for the estimation of parameters in complex models that involve economic or financial applications, where the identification of particular parameters is difficult, Bayesian methods may provide a useful synthesis between estimation and calibration techniques. Then lastly, Bayesian VAR models are also extremely powerful forecasting tools that allow for the researcher to place more weight on the information that is provided by the lags of a particular variable.

In the case of a multivariate model there are usually a large number of parameters that would need to be estimated, which restricts the degrees of freedom that would be calculated as, where $d.o.f. = (K^2 \times p) + C$. Therefore, with a VAR(4) that has five variables and a constant, we would need to provide estimates for 105 parameters. When using quarterly data, this would imply that we would need at least 27 years of data, when using classical techniques for parameter estimation. In addition, without prior information or some form of parameter shrinkage, it is hard to obtain precise estimates for all the parameters in a large multivariate model that is applied to a dataset that has limited observations. These techniques are also usually preferred when looking to derive parameter estimates for a State-Space model, as one would then be able to treat all the unobserved variables and parameters as random variables.

The choice between frequentist and Bayesian methods depends on the researchers preferences and on the task at hand. In practice, convenience may also play an important role, since in certain situations frequentist methods are easier to deal with, and in other cases Bayesian methods are more convenient. It has to be kept in mind, however, that these approaches may not only produce

numerically different answers, but that their interpretation is fundamentally different, even when the estimates coincide numerically. Since Bayesian methods are frequently used in VAR analysis, it is essential that most researchers should at least have a basic understanding of this approach.

Although Bayesian methods often require extensive computations, they have become quite popular for VAR analysis more recently as the cost of computing has decreased dramatically over the last few decades. Moreover, new methods and algorithms have broadened the applicability of Bayesian methods to many new and interesting research problems. In what follows we will consider some of the basic concepts of Bayesian modelling. For those looking for a complete treatment of the application of these techniques, see Canova (2007), Gelman et al. (2013), Geweke (2005), Koop (2003), Koop, Poirier, and Tobias (2007), Lancaster (2004), Poirier (1995) or Zellner (1971).

1 The Bayesian Paradigm

The ideas behind the Bayesian approach differ fundamentally from the classical inference (or the frequentist approach). Broadly speaking, when making use of classical inference we condition on the existence of a parameter vector, say Θ, that governs the data generating process (DGP) from which the observed sample of data is thought to have been obtained. The objective is to infer the value of Θ from this sample. Whereas the data is distributed according to the properties of a random variable (i.e. they are stochastic), the parameter vector, Θ, contains deterministic point estimates. Probability statements in such a case would refer to the properties of the estimator of Θ in a repeated sampling exercise.

In contrast, when making use of a Bayesian analysis, the parameter vector Θ contains a number of random variables, while the data could generated by a stochastic or nonstochastic process. Bayesians are concerned with modelling the researcher's beliefs about Θ. These beliefs are expressed in the form of a probability distribution. The Bayesian probability concept is essentially a subjective probability statement that does not require a repeated sampling exercise. Moreover, the nature of the DGP is immaterial for any inference relating to the parameter of interest, because inference conditions on the observed data.

The researcher's subjective beliefs about the value of the parameter vector, Θ, that are derived before they have inspected the data are summarized in the form of a prior probability distribution (or prior for Θ). This prior information is formally combined with the information contained in the data, as captured by the likelihood function of the model, to form the posterior probability distribution (or simply posterior) for Θ. This posterior conveys everything the researcher knows about the model parameters after having looked at the data.

When making use of classical inference one would acknowledge that we do not know the DGP for a given set of variables, but we nevertheless evaluate the properties of these methods under the assumption that there exists a true parameter vector, Θ, that can be objectively defined. Under such conditions, we postulate a DGP and then conduct our analysis, as if this model structure including any distributional assumptions were correct. In contrast, Bayesians do not need to make *a priori* assumptions about the underlying DGP. However, their formal framework for deriving and evaluating the posterior requires users to articulate their prior beliefs in the form of a prior probability distribution. It also involves assuming a specific distribution (or family of distributions) for the data, when evaluating the likelihood.

1.1 Prior, Likelihood, Posterior

The Bayesian approach treats the data, $y_t = \{y_1, \ldots, y_T\}$, as given and the parameter of interest, Θ, as unknown. Inference about Θ is conditional on the data. Prior information on Θ is assumed to be available in the form of a density.

Suppose the prior information is summarised in the prior probability density function (pdf) $g(\Theta)$ and the pdf for the underlying data generating process, conditional on a particular value of the parameter Θ is $f(y|\Theta)$. The latter function is algebraically identical to the likelihood function $\ell(\Theta|y)$. The two types of information are combined by applying Bayes' theorem, which states that the joint density is

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

where $f(\Theta, y)$ is the joint density, $g(\Theta|y)$ is the conditional probability and $f(y)$ is the marginal probability. Similarly, we could write $f(\Theta, y) = f(y|\Theta)g(\Theta)$. After equating the conditions for the joint density, we arrive at the conditional probabilities:

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

Here $f(y)$ denotes the unconditional sample density which is just a normalising constant for a given sample y_t and does not depend on Θ. Scaling the posterior by $f(y)$ ensures that the posterior density integrates to 1. In other words, the joint sample and prior information can be summarised by a function that is proportional to the likelihood function times the prior density $g(\Theta)$,

$$g(\Theta|y) \propto f(y|\Theta)g(\Theta) = \ell(y|\Theta)g(\Theta) \tag{1.1}$$

where $\propto$ in equation (1.1) indicates that the right-hand side is proportional to the left-hand side. The conditional density $g(\Theta|y)$ is the posterior pdf. It contains all the information that we have on the parameter vector Θ after having updated our prior views by looking at the data. Thus, the posterior pdf is the basis for estimation and inference as it summarises our beliefs about Θ, given our prior belief and the results of the likelihood function.

The computation of the posterior can be extremely intensive as it usually involves taking the product of distributions. Thereafter, the joint density is sampled by a subsequent algorithm to derive the individual parameter estimates. To illustrate what this would imply in a model that contains a single parameter, Figure 1 contains an example of a prior, which has a mean of 1.5 and is associated with a reasonable amount of uncertainty since the density is reasonably flat. The likelihood function is associated with higher degrees of certainty and is centred around a mean value of 2.3. This implies that the posterior, which summarises information about the prior and the likelihood function has a mean of 2.1, which is closer to the likelihood function as it would associated with more certainty.

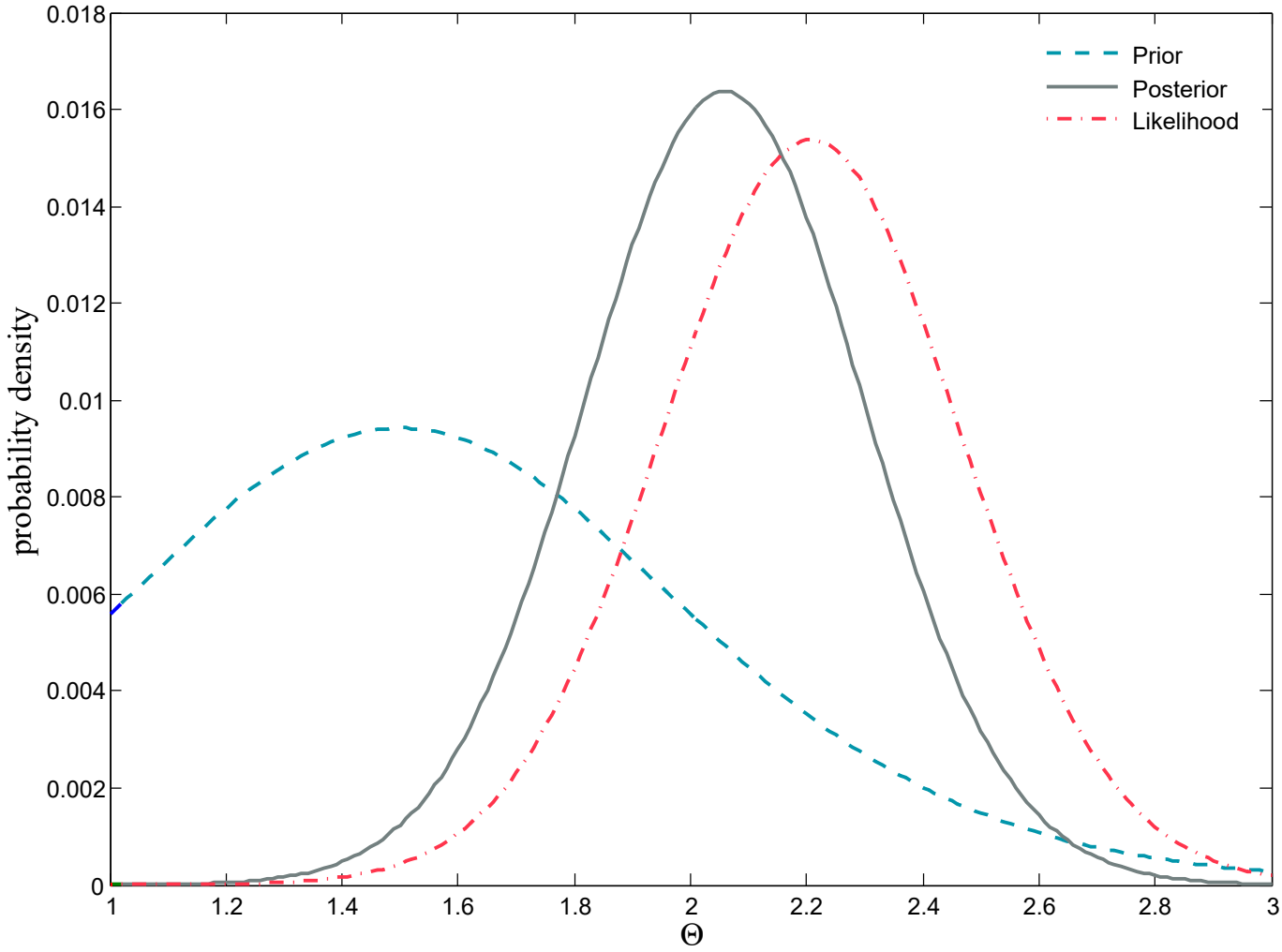

Figure 1: **Combining density functions to derive the posterior**

Note that 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. This inference is in many ways more useful than the classical case where we would traditionally calculate the probability of observing a single point estimate.

2 Priors for Reduced-Form VAR Parameters

In Bayesian analysis an important issue is the specification of the prior for the parameters of interest. Often a prior is specified that simplifies the posterior analysis. In particular, it is convenient to specify the prior such that the posterior is from a known family of distributions. If the prior is from the same distribution family as the likelihood function, then it is called a natural conjugate prior and the posterior will also have a distribution from the same family. 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, it is still necessary to specify at least some of the parameters of the prior distribution. This task is often made easier by imposing additional structure on the prior, reducing the number of parameters to be chosen to a handful of so-called hyperparameters. To explain this concept, let γ denote the vector of hyperparameters such that $g(\Theta) = g\gamma(\Theta)$. Often γ is chosen such that the implied VAR model yields accurate out-of-sample forecasts (see Doan, Litterman, and Sims (1984), Litterman (1986)). Alternatively, Bańbura, Giannone, and Reichlin (2010) suggest to choose these hyperparameters based on the in-sample fit of the model. Yet another proposal for choosing the hyperparameters was made by Giannone, Reichlin,

and Primiceri (2015). If one views the prior as being conditioned on the hyperparameters γ, $g\gamma(\Theta) = g(\Theta|\gamma)$, then the prior can be regarded as a hierarchical prior (see Koop (2003)). Suppose that the prior density for γ is $g(\gamma)$. Then the posterior is

$$g^{\star}(\gamma) \propto h(y|\gamma)g(\gamma),$$

where the sample density with respect to the hyperparameters is obtained as

$$h(y|\gamma) = \int f(y|\Theta,\gamma)g(\theta|\gamma)d\Theta$$

This expression is also known as the marginal likelihood because the parameters of interest, Θ, are integrated out. If an improper uniform prior, $g(\gamma) =$ constant, is specified, then the posterior of the hyperparameters is equal to the marginal likelihood, and it makes sense to choose the hyperparameters such that $h(y|\gamma)$ is maximised. Of course, strictly speaking, an improper prior does not qualify as a prior density because for an unbounded parameter space a constant prior does not integrate to one.

Giannone, Reichlin, and Primiceri (2015) stress two advantages of this approach. First, under certain conditions maximizing the marginal likelihood results in optimal out-of-sample forecasts (also see Geweke (2001) and Geweke and Whiteman (2006)). Second, they point out that their procedure also can be justified from a classical point of view.

2.1 The Minnesota Prior

Litterman (1986) and Doan, Litterman, and Sims (1984) propose a specific Gaussian prior for the parameters of a VAR model that is often referred to as the Minnesota prior or the Litterman prior. The original proposal shrinks the VAR estimates toward a multivariate random walk model. This practice has been found useful in forecasting many persistent economic time series. The proposal is to specify the prior mean and covariance matrix as follows:

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

where λ is the prior standard deviation of $a_{ii,1}$ and $0 < \theta < 1$ controls the relative tightness of the prior variance in the other lags in a given equation compared to the own lags (with a smaller θ increasing the relative tightness of the other lags). Note that σ_i^2 is the i^{th} diagonal element of Σ_u.

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

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

Here the numbers in parentheses are the prior standard deviations. 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 standard deviations for the intercept terms are set to infinity to capture our ignorance about the actual values of these parameters. Also, the prior distribution imposes independence across the parameters.

The crucial advantage of the Minnesota prior is that it reduces the problem of specifying a high-dimensional prior distribution to one of selecting two parameters by imposing additional structure on the prior. In specifying the Minnesota prior, the researcher has to choose only the two hyperparameters λ and θ. The parameter λ controls the overall prior variance of all VAR coefficients, whereas θ controls the tightness of the variances of the coefficients of lagged variables other than the dependent variable in a given equation. Roughly speaking, θ specifies the fraction of the prior standard deviation λ attached to the coefficients of other lagged variables. A value of θ close to one implies that all coefficients of lag 1 have about the same prior variance except for a scaling factor intended to capture differences in the variability of each variable. For example, Litterman (1986) finds that $\theta = 0.3$ and $\lambda = 0.2$ works well when using a Bayesian VAR model for forecasting U.S. macroeconomic variables. For given θ, the shrinkage is determined by λ. Therefore λ is often referred to as the shrinkage parameter. A smaller λ implies a stronger shrinkage towards the prior mean.[2]

If the above model were to represent a classical VAR and if the estimated value of the coefficient $a_{12,2}$ has a large value and a large standard error, then despite the large standard error, the large coefficient will still have a significant effect on the impulse response functions, forecasts, variance decompositions, etc. This would be most unfortunate and in contrast with these results, the BVAR could possibly ensure the effect of such a coefficient is very 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 structure of the prior ensures that this coefficient value shrinks to zero, given the structure of the prior.

There are also a number of other practical problems that have to be addressed in working with the Minnesota prior. For example, the assumption of a known Σ_u is unrealistic in practice. In a strict Bayesian approach, a prior pdf has to be specified for the elements of Σ_u and as such a simple alternative is to replace Σ_u by its LS estimator or its ML estimator. Of course one potential disadvantage is that Σ_u is not treated as a purely unknown parameter, as we replace the unknown values with an estimate that was derived outside of the model, thus we ignore much of the uncertainty that would have been generated during the estimation procedure. This approach can be improved upon by specifying a prior for Σ_u. As before, the prior distribution of the innovation covariance matrix must satisfy the constraint that in the reduced-form, Σ_u must be positive definite. Such models would traditionally make use of a natural conjugate Gaussian-Inverse Wishart Prior or an Independent Gaussian-Inverse Wishart Prior.

Another potential disadvantages of the Minnesota prior is that even if all variables have stochastic trends, it is not clear that shrinking towards a multivariate random walk as in the Minnesota prior is optimal because some of the variables may have cointegrated relationships. This approach can be rationalized on the grounds that exact unit roots are events of probability zero in standard Bayesian analysis. Hence, there is no reason to pay special attention to cointegration relations from a Bayesian point of view. Nevertheless the importance of the concept of cointegration and of VECMs in frequentist analysis has prompted some Bayesians to develop alternative priors that explicitly refer to the parameters of the VECM form of the VAR.

It is also worth noting that shrinking the parameters of models toward a random walk is only plausible for time series variables with stochastic trends. When working with stationary variables, the VAR parameters may be shrunk towards zero instead, as proposed by Lutkepohl (1991) and

implemented, for example, in Baumeister and Kilian (2012). In that case, mean-adjusting the data before fitting a VAR model may be useful to avoid having to specify a prior for the intercept term.

Many other modifications of the Minnesota prior have been proposed, depending on the needs of the researcher (see, for example, Kadiyala and Karlsson (1997), Sims and Zha (1998), Waggoner and Zha (2003), Bańbura, Giannone, and Reichlin (2010), Karlsson (2012)). The main advantage of the Minnesota prior is that it results in a simple analytically tractable normal posterior distribution, which explains why it has remained popular over the years, despite some disadvantages. However, there are alternatives, where by way of example, Sims and Zha (1998) proposed imposing prior restrictions on the structural VAR parameters rather than the reduced-form VAR parameters.

2.2 Illustration

To illustrate the use of the Minnesota prior, we consider a model including quarterly U.S. GDP deflator inflation (π_t), the seasonally adjusted unemployment rate (une_t) and the yield on the 3-month Treasury bills (r_t) for the period 1953q1 - 2006q3, as used by Koop and Korobilis (2010). The time series are plotted in Figure 2. All three series exhibit considerable persistence, so using the Minnesota prior with shrinkage to a random walk would appear to be reasonable.

We consider a VAR(4) model with intercept and impose a conventional Minnesota prior. Following the example of some earlier studies, the unknown error variances are replaced by estimates obtained from fitting univariate AR(4) models to the individual model variables. No estimates of the error covariances are required for the specification of the prior. In constructing the posterior, the error covariance matrix is treated as known and replaced by its LS estimate.

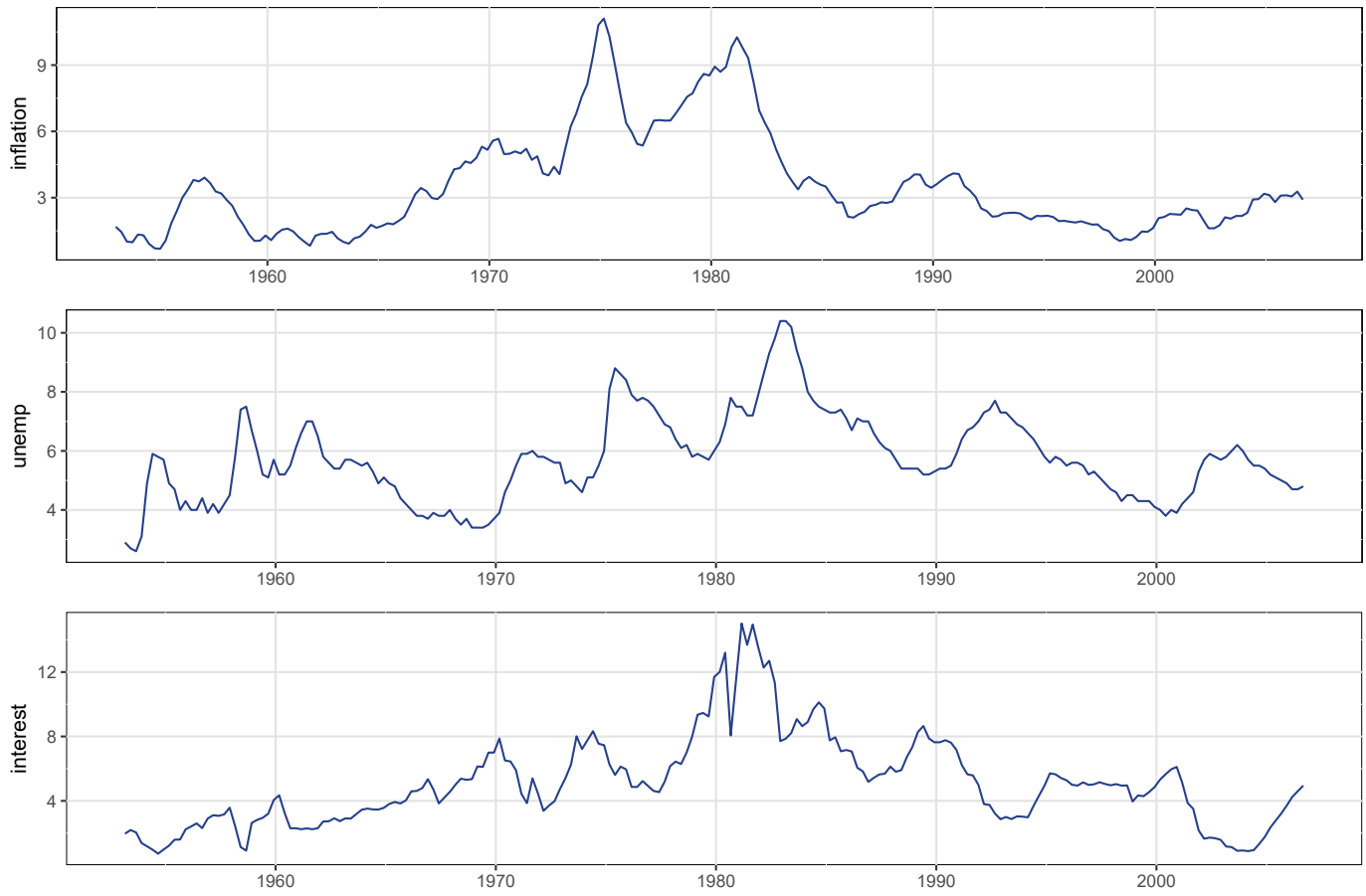

Figure 2: **Quarterly U.S. data**

Figure 3 illustrates the impact of various shocks on the posterior density of the structural impulse responses. The structural responses are obtained by imposing a recursive structure on the impact multiplier matrix with the variables ordered as $y_t = (\pi_t, une_t, r_t)'$. In this case, 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. Note the responses of inflation to an unexpected increase in the interest rate in the bottom left-hand corner. This would suggest that a contractionary monetary policy shock results in a decline in inflation (although the confidence intervals are relatively large), which would imply that the *price-puzzle* is not present.

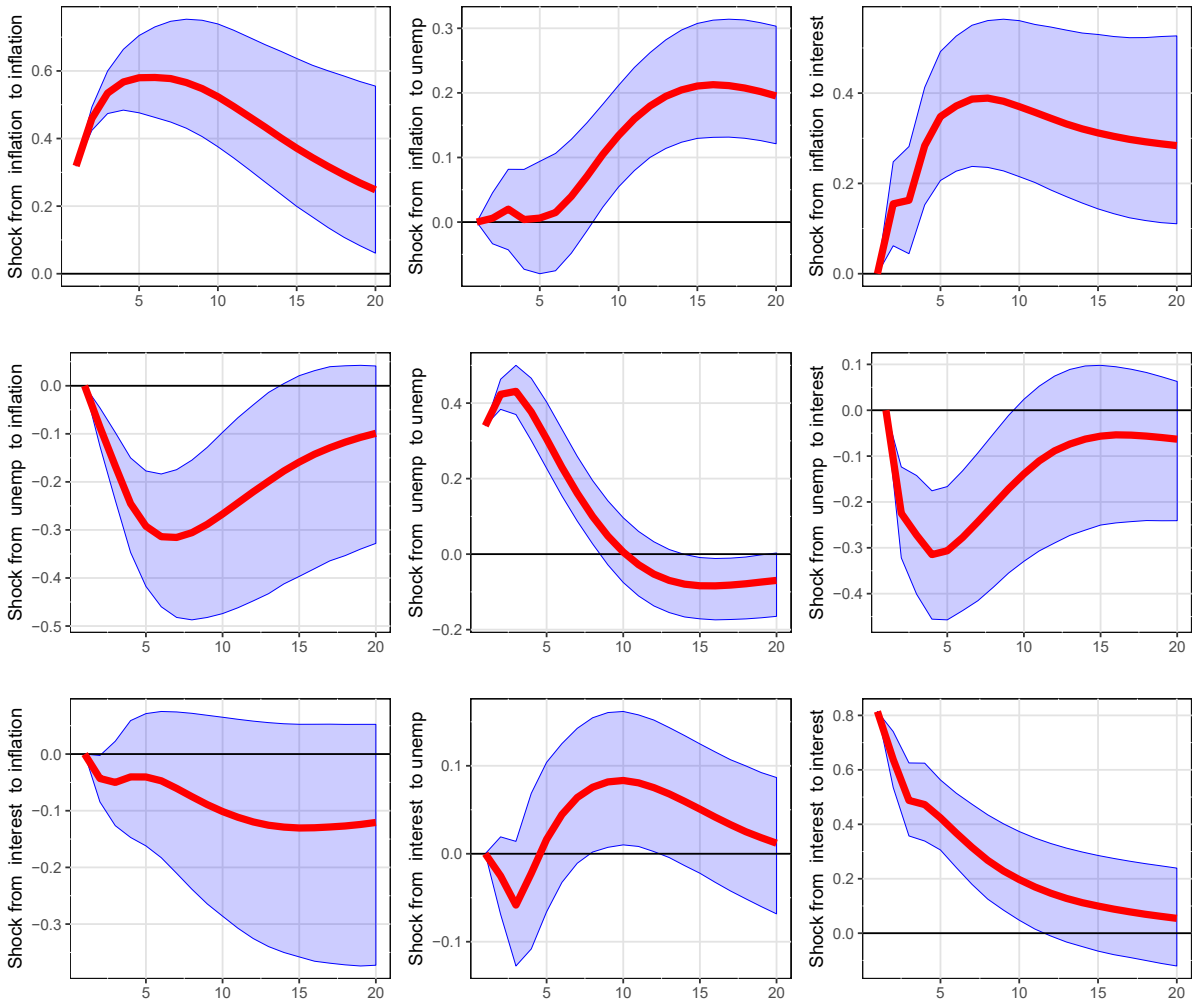

Figure 3: **Impulse response functions**

Following common practice, the figure plots the 10%, 50%, and 90% quantiles of the draws from the posterior distributions of the individual impulse response coefficients.

3 Models with sign restrictions

While the standard approach for identifying structural shocks in a VAR model would usually involve the application of short-run or long-run restrictions as has been applied in Sims (1980) and Blanchard and Quah (1989), such restrictions may be inconsistent with a number of theoretical structures (Canova and Pina 2005). Faust (1998), Canova and Nicolo (2002), and Uhlig (2005) present an alternative to this methodology, where they make use of prior beliefs about the signs of the impact of certain shocks derived from theoretical models to identify the structural shocks. In what

follows we consider the use of these methods, as implemented in the Uhlig (2005) rejection method, Uhlig (2005) penalty function method, Rubio-Ramìrez, Waggoner, and Zha (2010) rejection method, and Fry and Pagan (2011) median target method.

3.1 Identification with sign restrictions

To illustrate the idea behind recovering the structural shocks with the aid of sign restrictions, consider the following reduced-form VAR(1) model with k endogenous variables of the form:

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

where y_t is an $k \times 1$ vector of variables, A is an $k \times k$ matrix of coefficients, and u_t is a set of errors with mean zero, zero autocorrelation, and variance-covariance matrix

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

The reduced-form representation above summarises the sampling information in the data, where u_t is serially uncorrelated and orthogonal to the regressors in each equation. Note that in this specification u_t has no economic interpretation since the elements of u_t still might be correlated across equations.

While it is difficult to impose sign restrictions directly on the coefficient matrix of the model, it is easy to impose them ex-post on a set of orthogonalised impulse response functions. Thus, sign restrictions essentially explore the space of orthogonal decompositions of the shocks to see whether the responses conform with the imposed restrictions (Canova and Nicolo 2002).[3] In addition to making a choice about the signs of the responses, one has to specify for how long these restrictions apply after the impact of the shock. In theory, any length between the first period after the shock (only) and the entire length of response is possible.

The steps involved in recovering the structural shocks, given a set of sign restrictions, can be summarised as follows:

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. The Cholesky decomposition here is just a way to orthogonalise shocks rather than an identification strategy.
3. Calculate the resulting impulse responses from Step 2.
4. Randomly draw an orthogonal *impulse vector*, which is denoted 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.

Steps (4) and (5) would need further explanation, where 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, i.e. $||a|| = 1$ 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).

The type of decomposition differ slightly across the three implementations that we will consider. In the case of the two procedures suggested by Uhlig (2005), the generation of the impulse vector is based on a Givens rotation, whereas in the case of Rubio-Ramìrez, Waggoner, and Zha (2010) method is based on a QR-decomposition. As a result, the draw for a to get the impulse vector α differs as well. In the case of the Uhlig (2005) rejection method, one can directly draw from a $k \times 1$ vector in the unit sphere, whereas in the case of the Rubio-Ramìrez, Waggoner, and Zha (2010) rejection method, the draw is a $k \times 1$ vector from the standard normal distribution. In the case of the Uhlig (2005) penalty function method, a is based on a $(k-1) \times 1$ draw from a standard normal distribution and then projected into the R^n space using a stereographic projection.

Sign restrictions are, for the most part, only well defined from a Bayesian point of view Moon and Schorfheide (2012). 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 as discussed above.[4] One can then conduct inference based on the accepted draws and construct error bands similar to Sims and Zha (1999). While it might be desirable to have other priors, rather than an uninformative one, Baumeister and Hamilton (2015) show that the algorithms for identifying sign-restrictions only work for a very particular prior distribution, namely a uniform Haar prior and that standard informative priors will influence the posterior inference, even if the sample size is infinite. Thus, in the worst case scenario the researcher will simply analyse the prior distribution rather the posterior. A flat prior also has the advantage that standard (likelihood-based) specification tests, readily available in other packages, can be used and the researcher does not have to calculate, the marginal likelihood of the model. Lastly, a flat prior on the unit sphere is also appealing, since the results are unaffected by the Cholesky ordering in Step 2 (Uhlig 2005).

Step 5 involves constructing a suitable algorithm to check whether the impulse response functions have the appropriate sign or not. In the case of the Uhlig (2005) and Rubio-Ramìrez, Waggoner, and Zha (2010) rejection methods, the algorithm consists of a number of sub-draws to generate α for each posterior draw. The algorithm then checks whether the imposed sign restrictions are satisfied for each restricted response and each restricted period, starting off with the response of the shocked variable itself. In cases where the response of the shocked variable and α in the first period both have the wrong sign, the signs of the impulses and α are flipped, and the algorithm checks whether or not the restrictions are met or not. 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. If the draw is rejected, the algorithm tries another sub-draw for α from the same posterior draw to find a combination of the impulses and α that match the restrictions. If the algorithm cannot find a sub-draw for which the sign restrictions are met before reaching the maximum number of allowed sub-draws, it jumps to the next posterior draw after the maximum sub-draws.

In the case of the Uhlig (2005) penalty function, the algorithm is not based on the acceptance and rejection of sub-draws, as it seeks to find an impulse vector a which comes as close as possible to satisfying the imposed sign restrictions by minimising a function that penalises sign restriction violations. Let J be the total number of sign restrictions and N the total number of response periods for which the restrictions apply. The impulse vector here is the vector α which minimises the total penalty $\Psi(\alpha)$ for all constrained responses $j \in J$ at all constrained response periods $n \in N$. In order to treat all failures across the impulse responses symmetrically, the penalty function has to be adjusted for the scale of the variables and the sign of the restrictions. To treat the signs equally, let $l_j = -1$ if the sign of the restriction is positive and $l_j = 1$ if the restriction is negative. Scaling the variables is done by taking the standard error of the first differences σ_j of the variables as in Uhlig (2005).

Let $r_{j,\alpha}(n)$ be the response of j at response step n to the impulse vector α, then the minimisation problem can be written as

$$\min_{\alpha} \Psi(\alpha) = \sum_{j \in J} \sum_{n \in N} b \cdot f\left(l_j \frac{r_{j,\alpha}(n)}{\sigma_j}\right)$$

where b is a penalty depending on $f(\cdot)$ such that

$$b = \begin{cases} 1 & \text{if } f\left(l_j \frac{r_{j,\alpha}(n)}{\sigma_j}\right) \leq 0 \\ \text{penalty} & \text{if } f\left(l_j \frac{r_{j,\alpha}(n)}{\sigma_j}\right) > 0 \end{cases}$$

and *penalty* is a scalar ≥ 0 set by the user. While the specifications of $\Psi(\alpha)$ and b have no theoretical justification, Uhlig (2005) provides some discussion as to why the above forms are more suitable than others.

The next step is to find the impulse vector that minimises the total penalty for the restricted variables at all restricted horizons. Minimisation over the unit sphere of the above function is done as follows. First, we can follow Doan (2011) and parametrise the unit sphere in the k-space by randomly drawing a $k-1$ vector from a standard-Normal distribution and map the draw onto the k unit sphere using a stereographic projection.

There are two important features of sign restrictions, the user should be aware of. First, all three procedures that are implemented in what follows only allow for partial identification of the model. Hence, only one shock can be fully identified. While the rejection methods by Uhlig (2005) and Rubio-Ramìrez, Waggoner, and Zha (2010) can handle multiple shocks in theory, the penalty function approach, in its basic version, cannot. Mountford and Uhlig (2009) provide an extension of the penalty function approach that is able to cope with multiple shocks simultaneously. Second, the methods discussed here (and especially the rejection methods), are part of a class of so-called set identified models (Moon and Schorfheide 2012). This means that the parameters of interest are defined by a set of restrictions that do not point-identify them. Thus, the impulse responses of the models, at least in some cases, can only be bounded. The implications of set-identification for inference and why partial identification is not necessarily a bad thing and is discussed in latter sections.

3.2 Illustration

The original question in the paper of Uhlig (2005) was to analyse the effect of an unanticipated monetary policy shock, i.e an increase of one standard deviation in the United States (U.S.) Federal Reserve System (FED) short-term interest rate on real GDP, and the so-called "*price puzzle*", i.e. an increase in domestic inflation in response to a contractionary interest rate shock which is at the odds with economic theory. The data set contains monthly 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 1965:1 to 2003:12. Based on standard DSGE models, Uhlig (2005) suggested that an unanticipated innovation in the FED's policy rate

- does not decrease the FED's 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

Thus, Uhlig (2005) uses four restrictions in total to identify a monetary policy shock. The 1st and the 6th variable of the model (real GDP and total reserves) remain unrestricted. These sign restrictions that are used to identify a monetary policy shock in the model may be summarised as follows.

Table 1. Sign restrictions of the model

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

Given the ordering of the variables, with y_t appearing first, followed by π_t, p_t, i_t, rnb_t, and res_t, one can specify these restrictions (on the 2nd, 3rd, 4th, and 5th variable).

3.3 The Uhlig (2005) rejection method

The model is constructed as per the settings in Uhlig (2005), where we make use of 12 lags and no constant in the model. To generate the impulse response functions we make use of 60 steps. We use 200 draws from the posterior and 200 sub-draws for each posterior draw to generate the impulse vectors and the candidate impulse responses to which the rejection algorithm will be applied.

The number of rejected draws provides a crude indicator for how “well” the model is specified. A large number rejected draws, implies that there are potentially many other models that fit the data better than the one described by the current set of variables and/or sign restrictions. 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, as has been provided in Figure 4

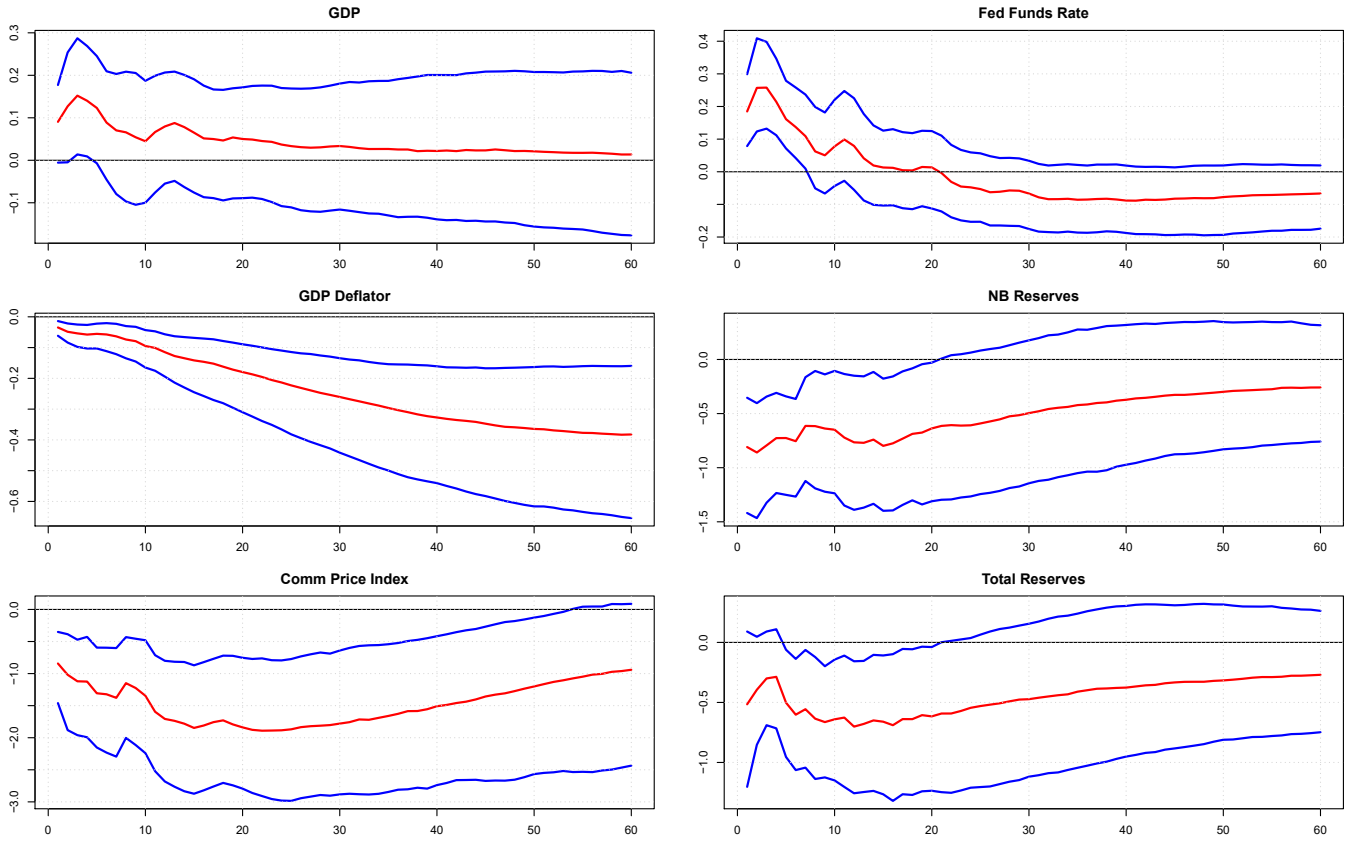

Figure 4: **Impulse response functions - Uhlig (2005) rejection method**

Note that in this way we obtain the desired result with regards to the obvious direction of the impulse response function. In the same way, one can plot the FEVDs of the model (measured in percentage points) for the shock of interest, which are displayed in Figure 5.[5]

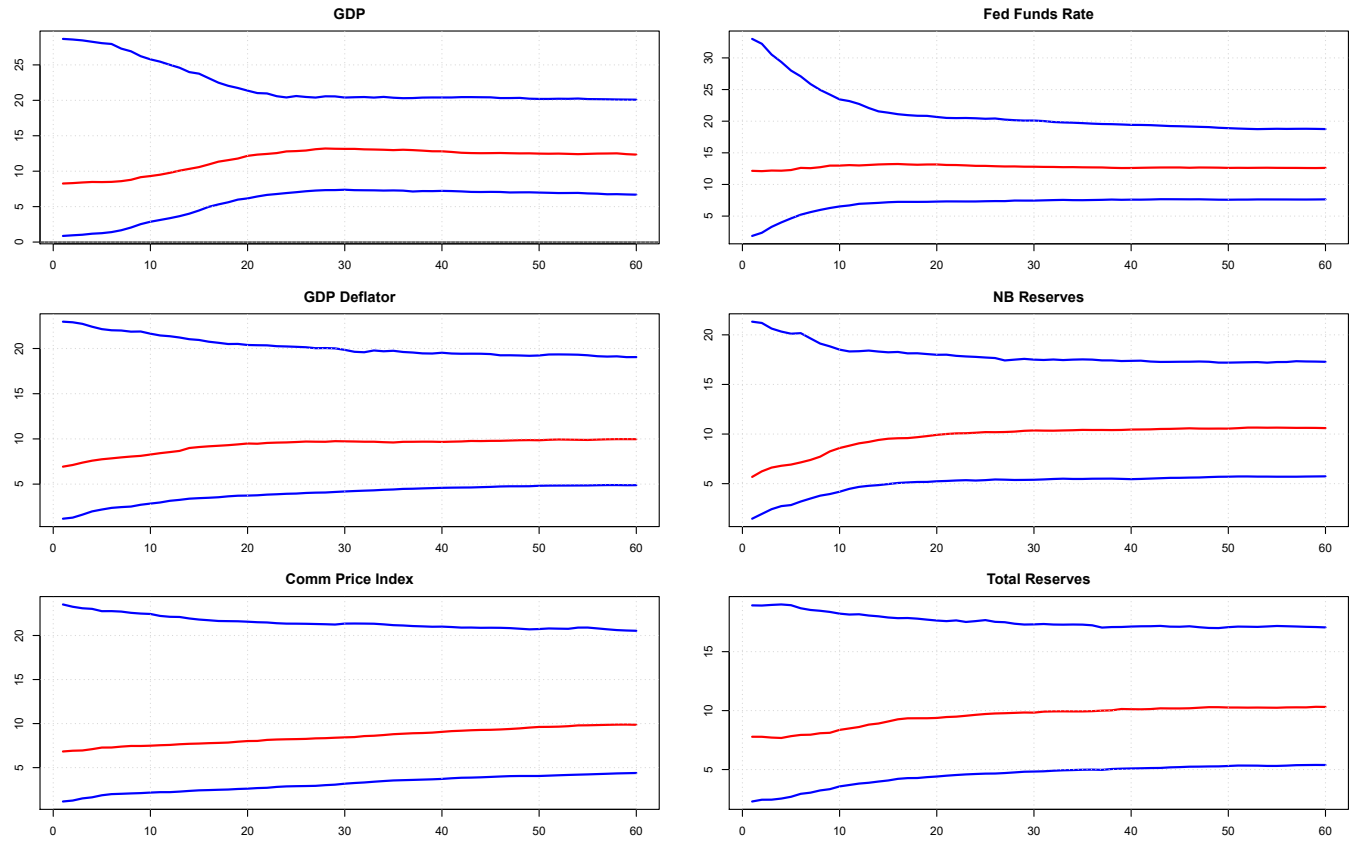

Figure 5: **Forecast error variance decomposition - Uhlig (2005) rejection method**

Then lastly, we are also able to show the implied shocks that may be derived from a particular variable in the model, where in Figure 6 we display the interest rate shocks.

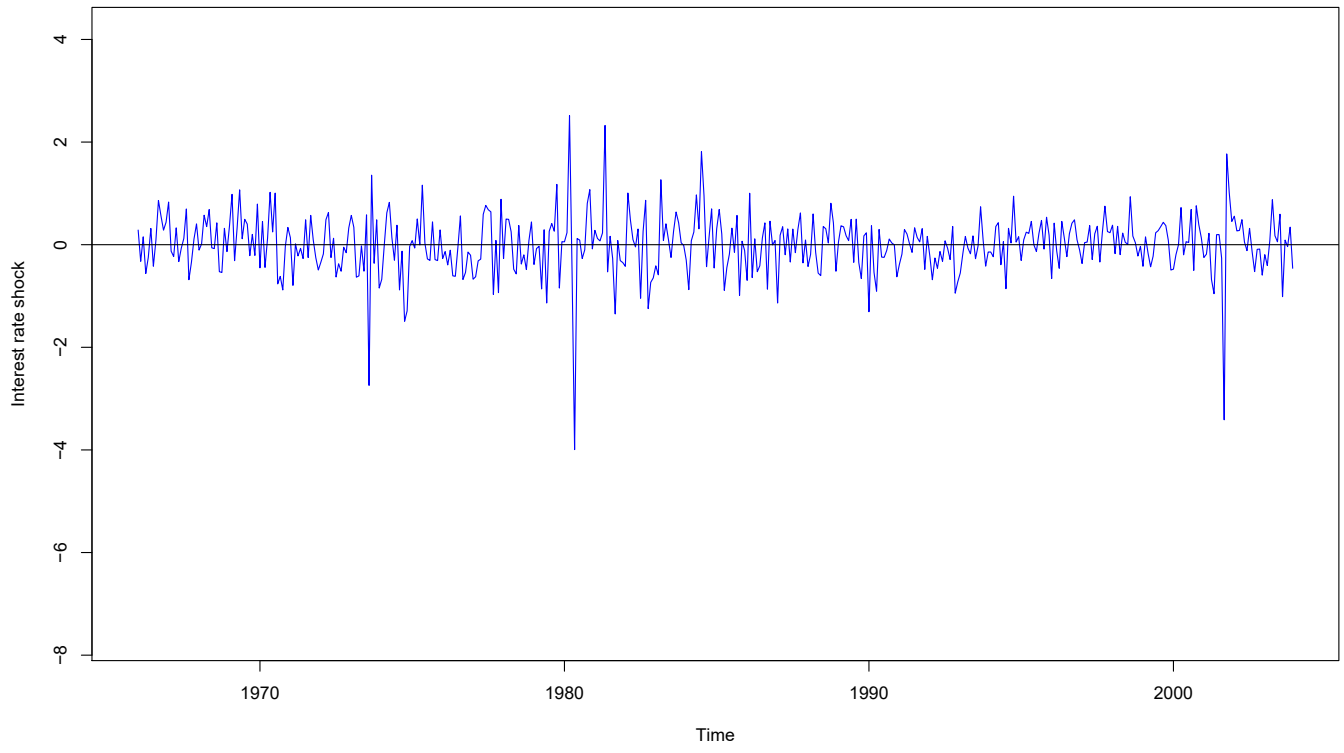

Figure 6: **Interest rate shock - Uhlig (2005) rejection method**

3.4 The Rubio-Ramirez et al. (2010) rejection method

When using the above parameterisation of the model we derive the same results, where the impulse response function is shown below.

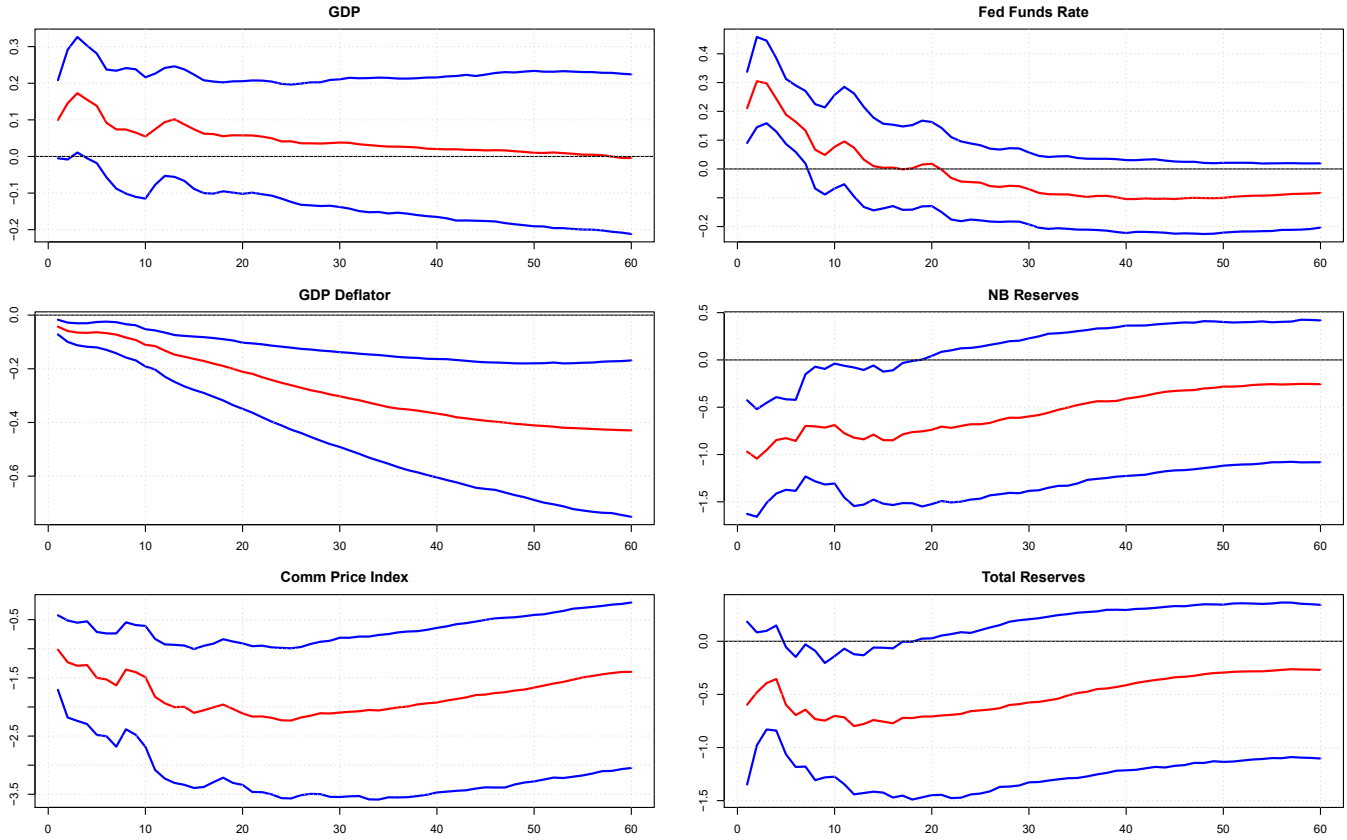

Figure 7: **Impulse response functions - Rubio-Ramirez et al. (2010) rejection method**

3.5 The Uhlig (2005) penalty function method

One shortcoming of the two rejection methods above is, that all impulse vectors satisfying the sign restrictions are considered to be equally likely (Uhlig 2005). Moreover, by construction, the rejection methods will only find impulse vectors that will exactly satisfy the sign restrictions. In some cases, there might be only very few impulse vectors which satisfy a full set of restrictions. A way to tackle this problem is with the use of the Uhlig (2005) penalty function method. As discussed previously, this condition minimises a criterion function that penalises for sign violations.

The penalty of $\Psi(\alpha)$ is used for the sign violations and since there are no sub-draws in the routine that can be used for inference, one needs a larger number of draws to generate the same number of desired acceptances. To ensure near-certain convergence of the minimisation problem, this procedure executes the minimisation algorithm twice for each draw of α. A draw here, gets rejected if either of the two minimisation runs do not converge, or if the two runs reach different optima, i.e. different values of $\Psi(\alpha)$ for the same draw of α. The critical value in this case is the (absolute) threshold value for the difference between the first and the second run of the minimisation routine above which the draw gets rejected.

In the current example, the penalty function needs 1003 draws to generate 1000 accepted draws. 3 posterior draws were rejected since the algorithm did not converge in either of the two runs or converged to two different optima. Hence, for 0.3 per cent of the draws the algorithm could not find a minimum value. This is an acceptable number and one can continue with the remainder of the

analysis. In those cases where there are a larger number of rejections due to non-convergence, the user may want to reconsider the model specification or the sign restrictions imposed on the responses.

Figure 8 shows the resulting impulse response function.[6]

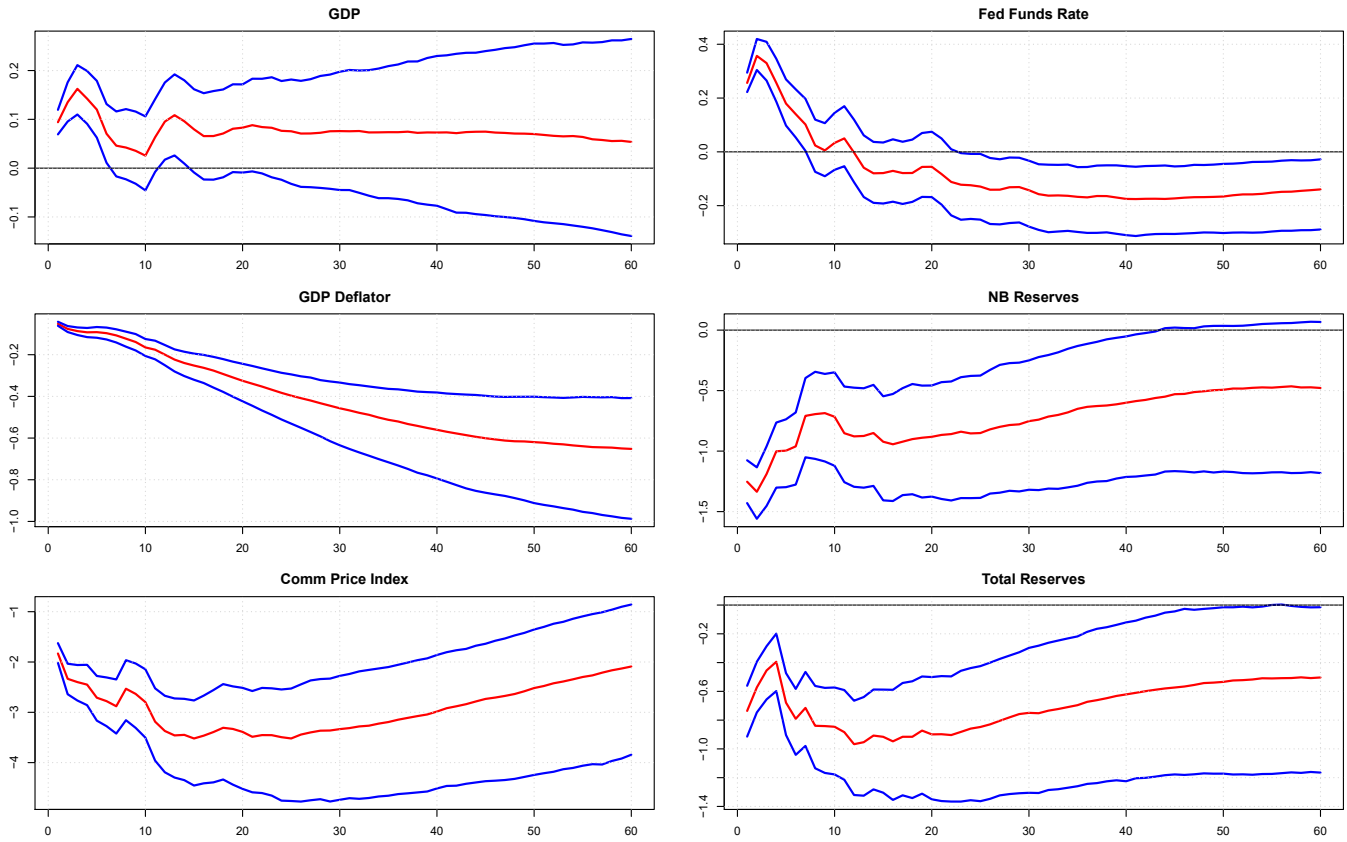

Figure 8: **Impulse response functions - Rubio-Ramirez et al. (2010) rejection method**

4 Critique and Issues

While sign restrictions provide a very appealing way of recovering the structural shocks from an economic point of view, several papers have pointed out a number of shortcomings of sign restrictions. Addressing these issues is subject to current research.

4.1 Multiple Shocks

What we have discussed are methods that allow for partial identification of the model. Partial identification is not necessarily a bad thing. The key idea behind sign restrictions is to characterise a shock through placing restrictions on the responses of some variables, but being agnostic about others. While it is possible to identify all shocks of the model, doing so by just using sign restrictions is inherently difficult. As shown by Fry and Pagan (2011) there are many cases in the literature, where researchers aim to identify all shocks in the model but fall short of doing so. One reason for this, is different shocks in the model might be characterised by the same set of restrictions. Considering the data set at hand, it is not too hard to imagine a model in which a GDP deflator shock is characterised by the same restrictions as a shock to the commodity price index. Another reason for underidentification is that researchers do not provide enough restrictions to uniquely identify all shocks in the model. 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.

4.2 Set Identification and Inference

As noted previously, models identified by sign restrictions are only set-identified. This means that, while sign restrictions solve the structural identification more elegantly than earlier methods, sign restrictions 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, each of which is equally consistent with both the observed data and the underlying restrictions (Moon et al. 2011). Hence, summarising the responses using e.g. the median response and conventional error bands to represent the spread of the responses displays the distribution across models and not, as intended, sampling uncertainty. The Uhlig (2005) and Rubio-Ramìrez, Waggoner, and Zha (2010) rejection methods are particularly prone to this "model identification" problem. While the question of whether there is *any* model for which the data can produce certain responses to shocks is an interesting one in itself, it is unsatisfactory in most cases.

There are several ways to deal with this problem. One way is to narrow down the set of admissible models to a singleton (Fry and Pagan 2011). This is what the Uhlig (2005) penalty function does by generating a "weighted sample" of all draws, including the ones that do not meet the imposed restrictions. Thus, if the results differ markedly across the three routines, the researcher should probably opt for the Uhlig (2005) penalty function specification.

A way to improve upon the results of the rejection methods is to use additional information and additional constraints. In general, the more restrictions one imposes, the "better" identified the model is, provided that the additional restrictions make sense. Uhlig (2005) and Arias, Rubio-Ramìrez, and Waggoner (2016) discuss ways of how to strengthen sign restrictions by imposing additional zero restrictions on selected impact responses. Imposing zero restrictions, however, brings back the problems discussed in the introduction, which Faust (1998), Canova and Nicolo (2002), and Uhlig (2005) set out to solve.

To demonstrate how additional restrictions can improve the model's results, consider the previous model with the four sign restrictions. As shown below, 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 Uhlig (2005) penalty function.

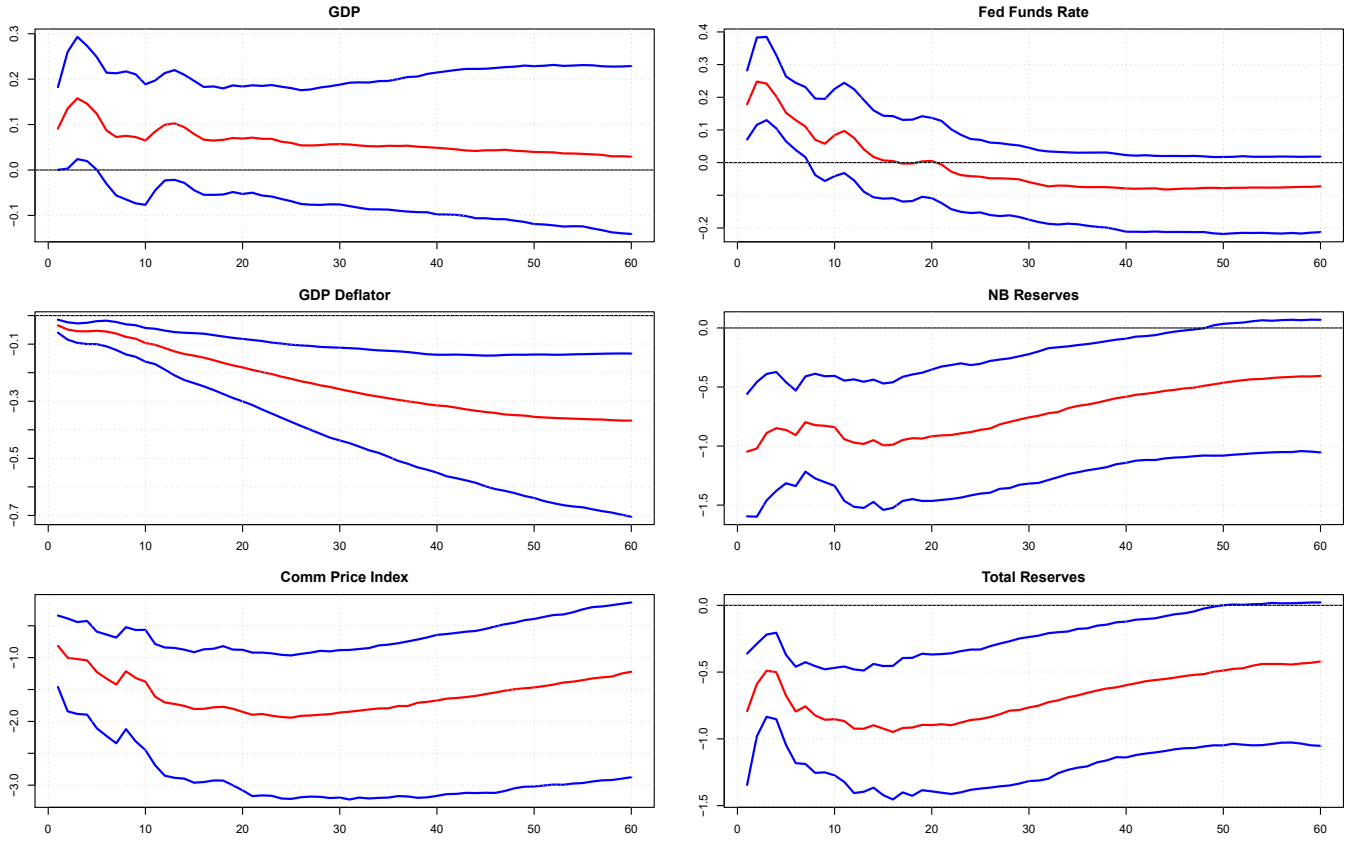

Figure 9: **Impulse response functions - Set identification**

In the same fashion, providing restrictions for the full set of shocks can solve the model identification problem as well. Canova and Paustian (2011) use simulated data from a DSGE model and show that sign restrictions recover the shocks of the model reasonably well, provided that enough restrictions are used and that all shocks are identified. This however, might not be feasible in many cases. Kilian and Murphy (2012) suggest an alternative way of achieving proper identification without imposing zero restrictions or identifying the full set of shocks by setting upper bounds for certain parameters of the model.

One way to analyse how significant the model identification problem is and how well the shocks in the model are identified, is Fry and Pagan (2011) Median-Target (MT) method. This method involves finding the *single* impulse vector α that produces impulse responses that are as close to the median responses as possible. The MT method is essentially a diagnostic device. Fry and Pagan (2011) show that strong differences between the MT impulse responses and the median responses indicate the standard model inference will be biased and misleading. The Fry and Pagan (2011) method finds the single best draw for α by minimisation is the sum of squared standardised gaps between the impulse responses given the test rotation and the sign restricted responses of the model that is tested.

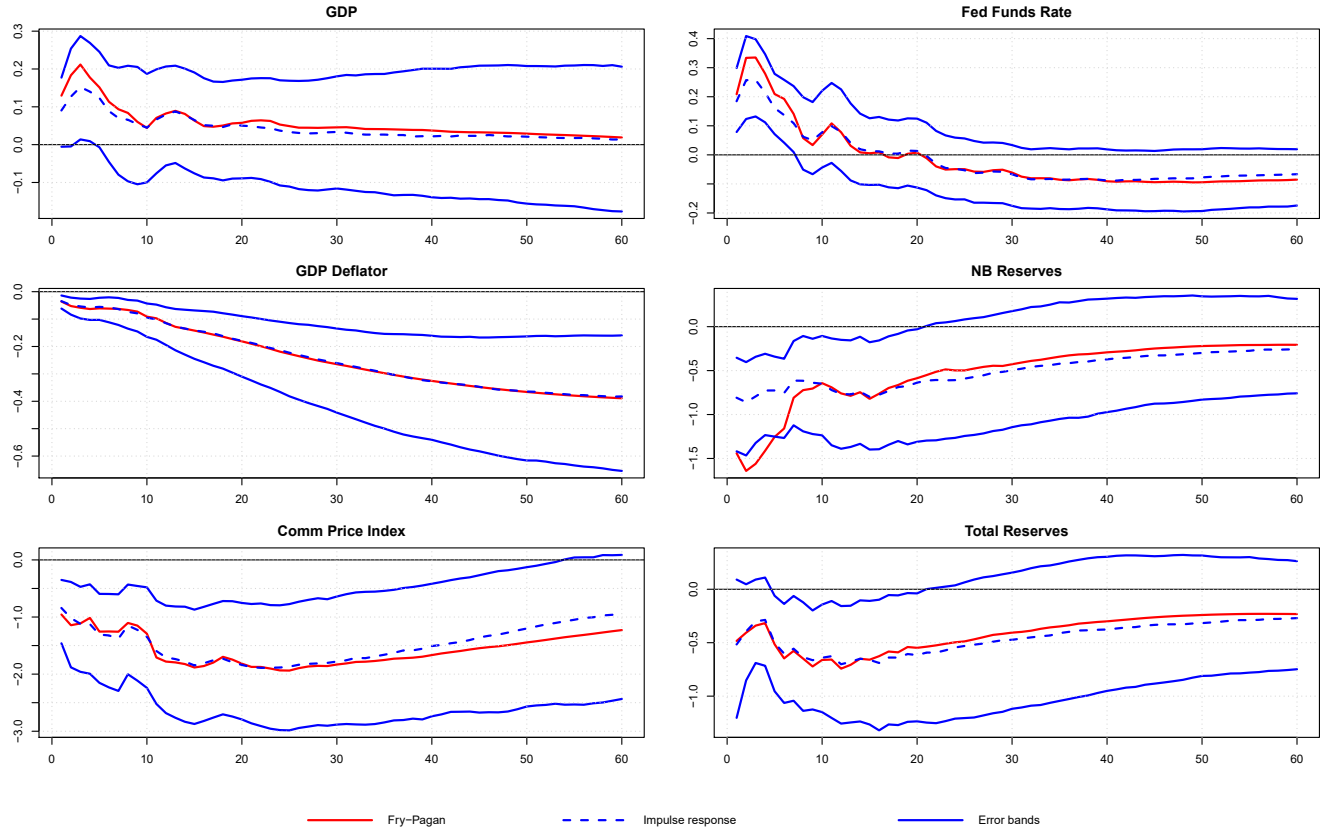

Figure 10: **Impulse response functions - Median-Target method**

Figure 10 shows the results of the impulse response function that makes use of this method. In this case the graph essentially compares the responses of Uhlig (2005) rejection method and the results of Fry and Pagan (2011) MT method. It's important to point out that the impulse responses of MT method do not have to lie within the error bands of the initial model, as for example, it is the case for the response of NB reserves to a monetary policy shock. The gap here, suggests that the median impulse response cannot be seen as a response of a single model and an alternative model should be considered. On the other hand, the remaining MT responses match the responses of Uhlig (2005) rejection method quite well in the long run. Something that one might consider evidence in favour of the current model specification.

5 Conclusion

This chapter considers the use of Bayesian VAR models, which continues to be a widely used approach in much applied work. It included a discussion of the Minnesota prior that may be applied to the reduced-form VAR model. This prior is able to apply a consistent method for imposing restrictions on potentially over-parameterised models, which results in a certain degree of parameter shrinkage. In addition, these models may be applied to non-stationary variables and have traditionally provided attractive out-of-sample forecasts for several economic variables. While alternative specifications for the variance-covariance matrix of the errors can be accommodated by the Bayesian framework, this generality usually increases the computational cost.

Thereafter, we considered the estimation of models that make use of sign restrictions. The models include the rejection and penalty function methods that were proposed by Uhlig (2005). In addition, we also considered the results of a model that makes use of the rejection method developed by Rubio-Ramìrez, Waggoner, and Zha (2010). After discussing pertinent details relating to the

implementation of the algorithms we were able to demonstrate the use of these routines. In addition, we also highlighted several shortcomings that the researcher should be aware of when implementing sign restrictions for these models.

6 References

Arias, Jonas, Juan F. Rubio-Ramìrez, and Daniel F. Waggoner. 2016. “Inference Based on Svars Identified with Sign and Zero Restrictions: Theory and Applications.” 2016 Meeting Papers 472. Society for Economic Dynamics.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin. 2010. “Large Bayesian Vector Auto Regressions.” *Journal of Applied Econometrics* 25 (1). Wiley Online Library: 71–92.

Baumeister, Christiane, and James D. Hamilton. 2015. “Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information.” *Econometrica* 83 (5): 1963–99.

Baumeister, Christiane, and Lutz Kilian. 2012. “Real-Time Forecasts of the Real Price of Oil.” *Journal of Business and Economic Statistics* 30 (2): 326–36.

Blanchard, Olivier J., and Danny Quah. 1989. “The Dynamic Effects of Aggregate Demand and Supply Disturbances.” *American Economic Review* 79 (4): 655–73.

Canova, Fabio. 2007. *Methods for Applied Macroeconomic Research*. Princeton: Princeton University Press.

Canova, Fabio, and Gianni De Nicolo. 2002. “Monetary Disturbances Matter for Business Fluctuations in the G-7.” *Journal of Monetary Economics* 49 (6): 1131–59.

Canova, Fabio, and Matthias Paustian. 2011. “Business Cycle Measurement with Some Theory.” *Journal of Monetary Economics* 58 (4): 345–61.

Canova, Fabio, and Joaquim Pires Pina. 2005. “New Trends in Macroeconomics.” In, edited by Claude Diebolt and Catherine Kyrtsou, 89–123. Berlin, Heidelberg: Springer Berlin Heidelberg.

Doan, Tom. 2011. “RATS Programs to Replicate Uhlig’s Var Identification Technique.” Estima.

Doan, Tom, Robert B. Litterman, and Christopher A. Sims. 1984. “Forecasting and Conditional Projection Using Realistic Prior Distributions.” *Econometric Reviews* 3 (1). Taylor & Francis: 1–100.

Faust, Jon. 1998. “The Robustness of Identified Var Conclusions About Money.” International Finance Discussion Papers 610. Board of Governors of the Federal Reserve System (U.S.).

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, and Juan F. Rubio-Ramìrez. 2010. “The Oxford Handbook of Applied Bayesian Analysis.” In, edited by Anthony O’Hagan and Mike West. Oxford: University of Pensilvania: Manuscript; Oxford University Press.

Fry, Renée, and Adrian Pagan. 2011. “Sign Restrictions in Structural Vector Autoregressions: A Critical Review.” *Journal of Economic Literature* 49 (4): 938–60.

Gelman, Andrew, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. 2013. *Bayesian Data Analysis*. Third. New York: Chapman; Hall/CRC.

Geweke, John. 2001. “Bayesian Econometrics and Forecasting.” *Journal of Econometrics* 100 (1): 11–15.

———. 2005. *Contemporary Bayesian Econometrics and Statistics*. New York, NY: John Wiley & Sons.

Geweke, John, and Charles Whiteman. 2006. “Bayesian Forecasting.” In, edited by G. Elliott, C. Granger, and A. Timmermann, 1st ed., 1:3–80. Elsevier.

Giannone, Domenico, Lucrezia Reichlin, and Giorgio Primiceri. 2015. “Prior Selection for Vector Autoregressions.” *The Review of Economics and Statistics* 2(97) (4). Elsevier: 436–51.

Kadiyala, K.R., and S. Karlsson. 1997. “Numerical Methods for Estimation and Inference in Bayesian VAR-Models.” *Journal of Applied Econometrics* 12 (2). John Wiley & Sons: 99–132.

Karlsson, S. 2012. “Forecasting with Bayesian Vector Autoregressions.” Working Paper 12/2012. Orebro University School of Business.

Kilian, Lutz, and Daniel Murphy. 2012. “Why Agonstic Sign Restrictions Are Not Enough: Understanding the Dynamics of Oil Market Var Models.” *Journal of the European Economic Association* 10 (5): 1166–88.

Koop, Gary. 2003. *Bayesian Econometrics*. Chichester: John Wiley & Sons.

Koop, Gary, and Dimitris Korobilis. 2010. “Bayesian Multivariate Time Series Methods for Empirical Macroeconomics.” *Foundations and Trends in Econometrics* 3: 267–358.

Koop, Gary, Dale J. Poirier, and Justin L. Tobias. 2007. *Bayesian Econometric Methods*. Edited by K. Abadir and J. Magnus & P.C.B. Phillips. Econometric Exercises 7. Cambrige: Cambridge Univ Press.

Lancaster, Anthony. 2004. *An Introduction to Modern Bayesian Econometrics*. Oxford: Blackwell.

Litterman, Robert B. 1986. “Forecasting with Bayesian Vector Autoregressions: Five Years of Experience.” *Journal of Business & Economic Statistics* 4 (1): 25–38.

Lutkepohl, H. 1991. *Introduction to Multiple Time Series Analysis*. Heidelberg: Springer-Verlang.

Moon, Hyungsik Roger, and Frank Schorfheide. 2012. “Bayesian and Frequentist Inference in Partially Identified Models.” *Econometrica* 80 (2): 755–82.

Moon, Hyungsik Roger, Frank Schorfheide, Eleonara Granziera, and Mihye Lee. 2011. “Inference for Vars Identified with Sign Restrictions.” Working Papers 11-20. Federal Reserve Bank of Philadelphia.

Mountford, Andrew, and Harald Uhlig. 2009. “What Are the Effects of Fiscal Policy Shocks?” *Journal of Applied Econometrics* 24 (6): 960–92.

Poirier, Dale J. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press.

Rubio-Ramìrez, Juan F., Daniel F. Waggoner, and Tao Zha. 2010. “Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference.” *Review of Economic Studies* 77 (2): 665–96.

Schorfheidede, Frank, and Marco Del Negro. 2011. “The Oxford Handbook of Bayesian Econometrics.” In, edited by J. Geweke, G. Koop, and H. van Dijk, 293–389. Oxford: Oxford University Press.

Sims, Christopher A. 1980. “Comparison of Interwar and Postwar Business Cycles.” *American Economic Review* 70 (2): 250–57.

Sims, Christopher A., and Tao Zha. 1998. “Bayesian Methods for Dynamic Multivariate Models.” *International Economic Review* 39 (4): 949–68.

———. 1999. “Error Bands for Impulse Responses.” *Econometrica* 67 (5): 1113–56.

Uhlig, Harald. 2005. “What Are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure.” *Journal of Monetary Economics* 52 (2): 381–419.

Waggoner, Daniel F., and Tao Zha. 2003. “A Gibbs Sampler for Structural Vector Autoregressions.” *Journal of Economic Dynamics and Control* 28 (2): 349–66.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York, NY: John Wiley & Sons.

---

1. The popular historian, Bill Bryson, has suggested that while Bayes was an extremely gifted mathematician, he was a poor preacher.↩
2. Rather than selecting λ and θ based on rules of thumb, one could instead treat these hyperparameters as endogenous and set them to the values that maximize the marginal likelihood, as proposed by Giannone, Reichlin, and Primiceri (2015).↩
3. Uhlig (2005) suggests that, to some extent, sign restrictions make explicit restrictions which are often used implicitly by the researcher in terms of what one considers to be a reasonable response to a shock.↩
4. Moon et al. (2011) apply sign restrictions using frequentist methods and show that the number of parameters to estimate can be very large and the topology of the identified set is quite complex.↩
5. Note that Uhlig (2005) does not show FEVDs as he makes use of forecast error revisions. Hence, these figures look slightly different to those that are contained in Uhlig (2005) for the analysis of variance.↩
6. Note that Uhlig (2005) only uses 100 accepted draws to construct the impulse responses of the penalty function in his paper. Due to the larger number of draws, the responses look slightly smoother than the ones shown in Uhlig (2005).↩