13 Identification by Sign Restrictions

The approach of using sign restrictions to identify structural VAR models was pioneered by Faust (1998), Canova and De Nicolo (2002), and Uhlig (2005). This approach has become increasingly popular in applied work as an alternative to traditional approaches to identification based on exclusion restrictions.

13.1 A Model of Demand and Supply

To understand why this approach is attractive from an economic point of view, consider a simple bivariate model of a goods market with a demand shock (w_t^{demand}) and a supply shock (w_t^{supply}). The observables are price (p_t) and quantity (q_t). Let us write the relation between the reduced-form residuals of a VAR model and the structural shocks as $u_t = B_0^{-1} w_t$, as in Chapter 8, where in our present example $u_t = (u_t^q, u_t^p)'$ and $w_t = (w_t^{\text{supply}}, w_t^{\text{demand}})'$. The implication of an exogenous shift in the demand or supply curve, respectively, for u_t^p and u_t^q depend on the slopes of the demand and supply curves, which may range from flat to vertical (see Figure 13.1).

A common approach of identifying the effects of demand and supply shocks by short-run exclusion restrictions relies on the short-run supply curve being vertical. In this case, demand shocks have no contemporaneous effect on quantity, which implies one exclusion restriction,

$$\begin{pmatrix} u_t^q \\ u_t^p \end{pmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} \begin{pmatrix} w_t^{\text{supply}} \\ w_t^{\text{demand}} \end{pmatrix}, \tag{13.1.1}$$

where asterisks denote unrestricted elements. This amounts to imposing that production does not respond within the impact period to a price increase triggered by a demand shock (see Figure 13.2). While this restriction may be reasonable in some contexts, as discussed in Chapter 8, in typical situations such recursive models are difficult to justify from an economic point of view.

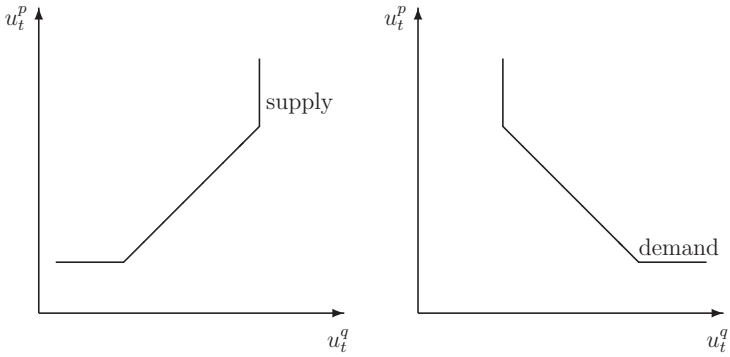

Figure 13.1. Demand and supply curves may have different slopes.

In general, all we know is that, under reasonable assumptions, a supply shock (represented by an exogenous shift to the right of the supply curve along the demand curve) will increase quantity and reduce the price, whereas a demand shock (represented as an exogenous shift to the right of the demand curve along the supply curve) will increase both price and quantity. These implications of economic theory relate to the sign of the responses of price and quantity to demand and supply shocks, respectively, and may be used for identification. Specifically, we may postulate that

$$\begin{pmatrix} u_t^q \\ u_t^p \end{pmatrix} = \begin{bmatrix} + & + \\ - & + \end{bmatrix} \begin{pmatrix} w_t^{\text{supply}} \\ w_t^{\text{demand}} \end{pmatrix}, \tag{13.1.2}$$

where $+$ and $-$ indicates a strictly positive and a strictly negative sign of the parameters in the structural impact multiplier matrix. Intuitively, knowing that positive supply shocks and positive demand shocks move the price in opposite directions, but quantity in the same direction helps us differentiate shifts of the demand curve from shifts of the supply curve (see Figure 13.3). If two shocks shared the same sign pattern, in contrast, one would be unable to identify them separately.

One key difference from the recursive model is that the parameters of the impact multiplier matrix in the sign-identified model are no longer point identified, but set identified. This means that even with an infinite amount of data we will only be able to bound the parameters of interest. As discussed later in this chapter, this fact greatly complicates estimation and inference in sign-identified VAR models.

In some studies, sign restrictions are represented as weak inequalities. The reason why weak inequalities are not permitted here is that identification would be lost if both weak inequalities were binding. For example, a solution for B_0^{-1}

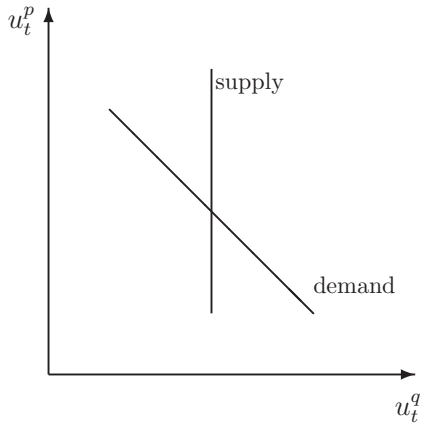

Figure 13.2. The case of a vertical supply curve.

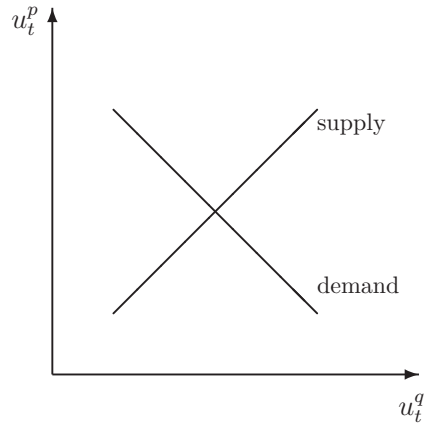

Figure 13.3. Standard demand and supply model.

of the form

$$\begin{bmatrix} + & + \\ 0 & 0 \end{bmatrix},$$

would be permitted under weak inequalities, but clearly is inadmissible because the two shocks would be observationally equivalent and B_0^{-1} would be rank deficient.

In contrast, it is possible in principle to achieve identification by judiciously combining weak and strict inequalities or by combining zero restrictions and strict inequalities. For example, the matrices

$$\begin{bmatrix} + & + \\ 0 & + \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} + & + \\ - & 0 \end{bmatrix},$$

would be admissible solutions for B_0^{-1} under alternative assumptions about the slopes of the demand and supply curves. The first of these matrices would be consistent with the pattern of demand and supply curves shown in Figure 13.4. The second matrix would be consistent with the pattern shown in Figure 13.5. Structural models that combine zero and sign restrictions will be discussed in more detail in Section 13.9.

It is sometimes argued that sign-identified models are more general than recursively identified models. Note, however, that the sign-identified model (13.1.2) does not nest the recursively identified model (13.1.1). While the sign-identified model relaxes the exclusion restriction in the recursively identified model, it does so at the cost of imposing sign restrictions on other parameters that were previously unrestricted. Thus, what we gain in generality in one dimension, we lose in the other dimensions. This means that the recursive

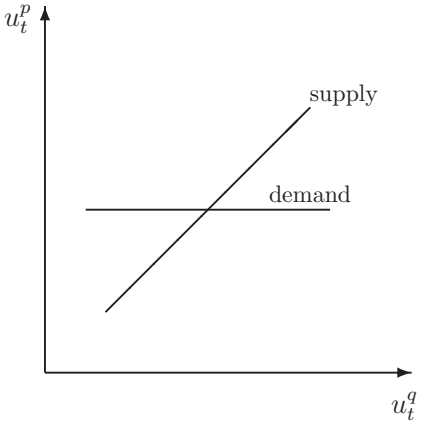

Figure 13.4. The case of a horizontal demand curve.

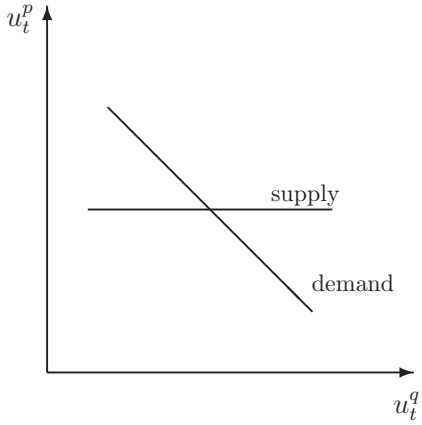

Figure 13.5. The case of a horizontal supply curve.

model is not nested in the sign-identified model. Hence, it is not possible to validate or invalidate the implications of recursively identified VAR models based on sign-identified models. Rather, these approaches are alternatives.

For now we focus on models identified by strict sign restrictions only. Although identification may be achieved by imposing a combination of exclusion and strict inequality restrictions, as discussed in Section 13.9, it is not possible to estimate sign-identified structural VAR models subject to weak inequalities. As shown later in this chapter, standard solution algorithms for sign-identified models are designed for strict inequalities only. All these algorithms have the property that if one were to impose a weak inequality on the response of the price to one shock, while imposing a strict inequality on the price response to the other shock, for example, the limiting case of zero is an event with probability measure zero.

One of the central questions is how to obtain numerical estimates of the structural model when only qualitative information is available about the model structure. It is useful to begin with the case of VAR models in which sign restrictions are imposed only on the impact responses of the observables to structural shocks. These responses correspond to the elements of the structural impact multiplier matrix, B_0^{-1}. Sign restrictions on the elements of B_0^{-1} are referred to as static sign restrictions.

13.2 How to Impose Static Sign Restrictions

Consider a structural vector autoregressive model

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t.$$

The variance-covariance matrix of the structural error term w_t is normalized such that

$$\mathbb{E}(w_t w_t') \equiv \Sigma_w = I_K.$$

Let $u_t = P\eta_t$, where u_t is the reduced-form VAR innovation and P is the lower-triangular Cholesky decomposition of Σ_u.[1] By construction, the shocks η_t are mutually uncorrelated and have unit variance. There is, of course, no reason for these shocks to correspond to economically interpretable structural shocks such as the demand and supply shocks in the bivariate model (13.1.2) above, but one can search for candidate solutions w_t^* for the unknown structural shocks w_t by constructing a large number of combinations of the shocks η_t of the form

$$w_t^* = Q'\eta_t,$$

where Q' is a square orthogonal matrix such that $Q'Q = QQ' = I_K$ and $u_t = PQQ'\eta_t = PQw_t^*$. Hence, each candidate solution w_t^* consists of uncorrelated shocks with unit variance. Whether any of these candidate solutions w_t^* is an admissible solution for the unknown structural shock w_t, given the vector of reduced-form parameters, depends on whether the implied structural impact multiplier matrix, PQ, satisfies the maintained sign restrictions on B_0^{-1}. We retain solutions that satisfy these sign restrictions and discard the remaining solutions. Repeating this procedure allows us to characterize the set of all structural models that are consistent with the maintained sign restrictions and the reduced-form parameters. More generally, knowledge of PQ allows the construction of all implied structural impulse response coefficients of interest from the estimates of the reduced-form slope parameters.

The ability to generate large numbers of candidate matrices Q from the set of all orthogonal matrices $\mathcal{O}$ thus is essential for the construction of sign-identified VAR models. In the following we denote the set of $K \times K$ orthogonal matrices by $\mathcal{O}(K)$, i.e.,

$$\mathcal{O}(K) \equiv \{Q \mid QQ' = I_K\}.$$

If the dimension of the matrices is not important or evident, we sometimes simply use $\mathcal{O}$ instead of $\mathcal{O}(K)$. There are two common approaches to constructing orthogonal matrices Q. One is based on Givens rotation matrices; the other is the Householder transformation approach.

[1] It should be noted that nothing hinges on P being the lower-triangular Cholesky decomposition. Any solution for P that satisfies $PP' = \Sigma_u$ will do as well.

13.2.1 Givens Rotation Matrices

Givens rotation matrices can be used to construct orthogonal matrices. In the bivariate model, Givens matrices have the form

$$Q(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix},$$

where ϕ lies between 0 and 2π. Each choice of $\phi \in [0, 2\pi]$ implies an orthogonal matrix $Q(\phi)$ and

$$\mathcal{O}(2) = \{Q(\phi) \,|\, \phi \in [0, 2\pi]\}.$$

In practice, one defines a finite-dimensional grid over the possible values of ϕ (or, alternatively, draws ϕ from a uniform distribution on $[0, 2\pi]$), computes the implied $Q(\phi)$ and the corresponding structural impact multiplier matrices $PQ(\phi)$, and retains only those solutions that agree with the maintained sign restrictions. This allows one to characterize the space of admissible structural VAR models, with each structural model defined as a set of structural impulse responses, conditional on the reduced-form estimate.

In the trivariate model, one way of forming an orthogonal matrix Q is to generate the product

$$Q(\phi_1, \phi_2, \phi_3) = Q_{12}(\phi_1) \times Q_{13}(\phi_2) \times Q_{23}(\phi_3),$$

of the Givens rotation matrices

$$Q_{12} = \begin{bmatrix} \cos\phi_1 & -\sin\phi_1 & 0 \\ \sin\phi_1 & \cos\phi_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; Q_{13} = \begin{bmatrix} \cos\phi_2 & 0 & -\sin\phi_2 \\ 0 & 1 & 0 \\ \sin\phi_2 & 0 & \cos\phi_2 \end{bmatrix},$$
$$Q_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_3 & -\sin\phi_3 \\ 0 & \sin\phi_3 & \cos\phi_3 \end{bmatrix},$$

where each ϕ_i, $i = 1, 2, 3$, lies between 0 and 2π. Given $\cos^2\phi_i + \sin^2\phi_i = 1$, it can be shown that $Q_{12}'Q_{12} = I_3$, $Q_{13}'Q_{13} = I_3$, and $Q_{23}'Q_{23} = I_3$. By construction, $Q(\phi_1, \phi_2, \phi_3)$ is an orthogonal matrix.

Canova and De Nicolo (2002) suggest defining a finite-dimensional grid of values for each ϕ_i between 0 and 2π, computing all implied $Q(\phi_1, \phi_2, \phi_3)$, and retaining only those solutions that yield a structural impact multiplier matrix $PQ(\phi_1, \phi_2, \phi_3)$ that agrees with the maintained sign restrictions.

Generalizations to $K > 3$ are possible, but it is clear that this algorithm becomes computationally burdensome for large-dimensional VAR models and may be computationally infeasible for large K, which explains why it is rarely used in practice.

13.2.2 The Householder Transformation

An alternative and more common approach for computing suitable impact multiplier matrices, proposed by Rubio-Ramírez, Waggoner, and Zha (2010), is an algorithm involving the QR decomposition (or QR factorization). Recall that any real square matrix W can be decomposed as $W = QR$, where Q is an orthogonal matrix (i.e., its columns are orthogonal unit vectors such that $QQ' = I$) and R is an upper-triangular matrix. If W is invertible, then the factorization will be unique, provided the diagonal elements of R are restricted to be positive. The QR factorization of a matrix may be computed using the Householder transformation discussed in Stewart (1980).

The algorithm suggested by Rubio-Ramírez, Waggoner, and Zha (2010) covers the space $\mathcal{O}(K)$ of $K \times K$ orthogonal matrices Q by drawing each column of a $K \times K$ matrix W at random from a $\mathcal{N}(0, I_K)$ distribution and applying the QR factorization to each draw W. They show that if we choose Q by a QR decomposition of W with the diagonal of the upper triangular matrix R normalized to be positive, then this amounts to drawing Q from a uniform distribution over the space of orthogonal matrices $\mathcal{O}(K)$. Drawing the elements of the matrix W independently from a $\mathcal{N}(0, 1)$ distribution will result in a nonsingular matrix with probability 1. This fact ensures that invertibility holds when applying this algorithm. The mathematical underpinnings of this approach follow from Stewart (1980).

The algorithm can be used for generating a large number of candidate solutions for B_0^{-1} as PQ, where P denotes the lower-triangular Cholesky decomposition of Σ_u and Q is obtained from a random draw for W. The matrix Q in the literature is often referred to as the rotation matrix. Fry and Pagan (2011) observe that this method is equivalent to using Givens matrices. Its main advantage is that it is more computationally efficient for large K.

This leaves the question of how to compute the QR decomposition in practice. Commonly used software provides built-in functions for this purpose such as the *qr* function in MATLAB. Note, however, that this function does not ensure positive diagonal elements of R. Hence, that normalization has to be performed on the output of the *qr* function by reversing all signs of each row of R corresponding to a negative diagonal element and adjusting Q accordingly to ensure that $W = QR$ holds. Since we are not interested in R, we may equivalently just reverse all signs in the i^{th} column of Q if the i^{th} diagonal element of R is negative. Note also that the sign normalization should not be performed on Q because that may violate the premise of uniform sampling from $\mathcal{O}(K)$. For example, if one were to standarize the main diagonal elements of Q instead of R to be positive, one would still obtain a unique QR decomposition. However, one would miss all elements of $\mathcal{O}(K)$ with negative diagonal elements.

Having obtained many candidate solutions for B_0^{-1}, one retains only those solutions that yield a structural impact multiplier matrix that agrees with the

maintained sign restrictions. More generally, knowledge of admissible B_0^{-1} matrices also allows the construction of the implied set of structural impulse responses from the reduced-form slope parameters. Any draw for B_0^{-1} that does not satisfy all sign restrictions on the impulse responses must be rejected. The simplest approach is to discard this draw and to generate another draw. This approach is computationally inefficient if all responses to a given shock have the wrong sign. Rubio-Ramírez, Waggoner, and Zha (2010) observe that in that case, instead of generating a new draw, one can simply change the sign of the column of Q in question, resulting in a new orthogonal matrix that satisfies the sign restrictions. Notice that multiplying a column of an orthogonal matrix Q by -1 results in another orthogonal matrix.

A substantial further computational gain may be achieved by checking each column of PQ for the sign pattern associated with a given shock, exploiting the fact that in sign-identified VAR models the ordering of the variables does not determine which shock is contained in which column of the structural impact multiplier matrix (see, e.g., Baumeister and Hamilton 2015a). Suppose we are interested in a bivariate model with sign restriction matrix

$$\begin{bmatrix} + & + \\ - & + \end{bmatrix},$$

where the first column refers to the impact effects of a supply shock and the second column refers to a demand shock. If we obtain a draw

$$\begin{bmatrix} 0.41 & 0.01 \\ 0.30 & -0.88 \end{bmatrix}$$

for B_0^{-1}, we can interpret the first column as representing the impact effects of the demand shock and the second column as representing the impact effects of the supply shock rather than discarding this draw simply because the signs of the first column do not match the sign restriction matrix. This approach is equivalent to taking an alternative draw for Q that flips the columns of this matrix.

13.2.3 The Ouliaris-Pagan Approach

Yet another approach for computing suitable structural impact multiplier matrices was proposed by Ouliaris and Pagan (2016) who focus on the problem of imposing static sign restrictions on the elements of B_0 rather than on the elements of the structural impact multiplier matrix. Let B_0 have a unit diagonal and Σ_w be a diagonal matrix with unrestricted main diagonal. Then, if Σ_u and the $K(K-1)/2$ elements of B_0 above the main diagonal are known, one can solve the system

$$\Sigma_u = B_0^{-1} \Sigma_w B_0^{-1\prime} \tag{13.2.1}$$

for the elements of B_0 below the main diagonal and for the diagonal elements of Σ_w. Building on this observation, Ouliaris and Pagan propose to draw the above-diagonal elements of B_0 at random such that possible sign restrictions on B_0 are preserved. They then solve for the remaining elements of B_0 and the diagonal elements of Σ_w. They retain the resulting candidate solution for B_0 if all sign restrictions on B_0 are satisfied. Otherwise it is discarded. The procedure is repeated many times to ensure that the entire admissible parameter space is represented.

The procedure for drawing the above-diagonal elements of B_0 is as follows. For the ij^{th} element of B_0, denoted $b_{ij,0}$, with $i < j$, a random variable ϕ is drawn from the uniform distribution

$$\begin{array}{lll} \mathcal{U}(-1,1) & \text{if} & b_{ij,0} \text{ is unrestricted,} \\ \mathcal{U}(0,1) & \text{if} & b_{ij,0} > 0, \\ \mathcal{U}(-1,0) & \text{if} & b_{ij,0} < 0. \end{array}$$

Then $b_{ij,0}$ is set to $\phi/(1-|\phi|)$. Independent draws are used for all elements $b_{ij,0}$ with $i < j$. Note that this algorithm, rather than sampling from the space of all possible orthogonal matrices Q, samples from the set of possible values for the elements of B_0, treating these elements as independent. By construction, $b_{ij,0}$ has infinite support, but the algorithm assigns more probability mass to values of $b_{ij,0}$ close to zero.

Given the above-diagonal elements of B_0 and an estimate of Σ_u, the below-diagonal elements of B_0 and the diagonal elements of Σ_w can be estimated by solving the system (13.2.1) with a nonlinear equation solver as discussed in Chapter 10. An alternative possibility is to use IV estimation for estimating the values of the below-diagonal elements. Given the recursive structure of the estimation problem, we can use LS estimation of the first equation,

$$y_{1t} + b_{12,0}y_{2t} + \cdots + b_{1K,0}y_{Kt} = \sum_{k=1}^{K}\sum_{i=1}^{p} b_{1k,i}y_{k,t-i} + w_{1t},$$

where the left-hand-side parameters are fixed. The estimation error $\widehat{w}_{1t}$ can then be used as an instrument for y_{1t} in estimating the second equation,

$$y_{2t} + b_{23,0}y_{3t} + \cdots + b_{2K,0}y_{Kt} = -b_{21,0}y_{1t} + \sum_{k=1}^{K}\sum_{i=1}^{p} b_{1k,i}y_{k,t-i} + w_{2t},$$

where again the left-hand-side parameters are given. More generally, the k^{th} equation,

$$\begin{aligned} &y_{kt} + b_{k,k+1,0}y_{k+1,t} + \cdots + b_{kK,0}y_{Kt} \\ &\quad = -b_{k1,0}y_{1t} - \cdots - b_{k,k-1,0}y_{k-1,t} + \sum_{k=1}^{K}\sum_{i=1}^{p} b_{1k,i}y_{k,t-i} + w_{kt}, \end{aligned}$$

can be estimated by using the residuals from the previous equations, $\widehat{w}_{jt}$, as the instrument for y_{jt}, $j = 1, \ldots, k-1$. If the estimates obtained in this way satisfy the sign restrictions, the implied estimate of B_0 is retained. Otherwise it is discarded.

As an example, consider a 3-dimensional VAR(1) model where no inequality restrictions are imposed on the above-diagonal elements of

$$B_0 = \begin{bmatrix} 1 & b_{12,0} & b_{13,0} \\ b_{21,0} & 1 & b_{23,0} \\ b_{31,0} & b_{32,0} & 1 \end{bmatrix}.$$

The first step of the procedure is to estimate the reduced form, $y_t = A_1 y_{t-1} + u_t$, by LS. We obtain an estimate $\widehat{\Sigma}_u$ of the reduced-form error covariance matrix. We then draw three independent numbers ϕ_1, ϕ_2, ϕ_3 from a uniform distribution $\mathcal{U}(-1, 1)$ and fix the above-diagonal elements of B_0 at $b^*_{12,0} = \phi_1/(1 - |\phi_1|)$, $b^*_{13,0} = \phi_2/(1 - |\phi_2|)$, and $b^*_{23,0} = \phi_3/(1 - |\phi_3|)$, respectively. Finally, we estimate $b_{21,0}$, $b_{31,0}$, and $b_{32,0}$ and the diagonal elements of Σ_w by solving (13.2.1) for the unknown parameters with Σ_u replaced by $\widehat{\Sigma}_u$.

Alternatively, estimates of the unknown parameters may be obtained by LS estimation of

$$y_{1t} + b^*_{12,0} y_{2t} + b^*_{13,0} y_{3t} = b_{11,1} y_{1,t-1} + b_{12,1} y_{2,t-1} + b_{13,1} y_{3,t-1} + w_{1t}.$$

The residuals of this regression then are used as the instrument for y_{1t} in the IV estimation of

$$y_{2t} + b^*_{23,0} y_{3t} = -b_{21,0} y_{1t} + b_{21,1} y_{1,t-1} + b_{22,1} y_{2,t-1} + b_{23,1} y_{3,t-1} + w_{2t}.$$

Finally,

$$y_{3t} = -b_{31,0} y_{1t} - b_{32,0} y_{2t} + b_{31,1} y_{1,t-1} + b_{32,1} y_{2,t-1} + b_{33,1} y_{3,t-1} + w_{3t}$$

is estimated by IV using the residuals from the previous two equations as instruments for y_{1t} and y_{2t}.

This method may also be adapted to allow the B_0 matrix to be restricted in other ways, provided that enough model parameters are fixed, so the remaining parameters can be solved for or estimated. For example, if the main diagonal of B_0 is unrestricted, Σ_w is the identity matrix, and Σ_u is known, the system $\Sigma_u = B_0^{-1} B_0^{-1\prime}$ can be solved for the elements on and below the main diagonal, as long as the above-diagonal elements of B_0 are fixed by assigning random draws to them.

13.3 Partially Identified VAR Models

Implicitly, the discussion thus far assumed that the model is fully identified in that all structural shocks are individually identified. This is not always the case.

A common situation in applied work is that the researcher only has knowledge of the signs of the responses to some shocks. A situation in which the number of identified shocks is less than K is known as a partially identified VAR model (see, e.g., Rubio-Ramírez, Waggoner, and Zha 2010; Inoue and Kilian 2013). Most common are situations in which only a single structural shock is of interest.[2]

For example, consider the following partially identified bivariate model. Suppose that structural shock w_{1t} is known to raise both observables on impact, whereas we have no a priori knowledge of the sign of the responses to structural shock w_{2t}:

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} = \begin{bmatrix} + & ? \\ + & ? \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}.$$

It may seem that in this case it would be enough to check the signs in the first column of the structural impact multiplier matrix. This is not the case if we insist that the sign pattern of the first shock must be distinct from that of the other shock. In that case, we still need to verify the signs of the responses to w_{2t} because a realization of PQ may give rise to a candidate solution

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} = \begin{bmatrix} + & + \\ + & + \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}.$$

Given that we can interchange the two candidate structural shocks in this case, it is not clear which of the two shocks we should focus on. Since the two shocks are orthogonal, they are distinct in the statistical sense, but they are not distinct in the economic sense. Simply choosing the first structural shock to be the demand shock would be arbitrary. Moreover, we know that this particular candidate solution is inconsistent with any reasonable economic model of this market because it implies the existence of two demand shocks but no supply shock. Hence, we need to verify that the signs in the second column are the complement of the signs in the first column, namely,

$$\begin{bmatrix} + & + \\ + & - \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} + & - \\ + & + \end{bmatrix}.$$

If they are not, this draw for PQ must be discarded.

Although in this simple example the requirement that the identified shock be distinguished by a unique sign pattern effectively implies a fully identified

[2] Set identification arises whenever inequality constraints are imposed. In microeconometrics it is common to refer to set-identified models interchangeably as partially identified models. In the present context, partial identification has a different meaning, and it is important to keep in mind that a model may be set identified without being partially identified and partially identified without being set identified. Nevertheless, there are some VAR studies that ignore this distinction and refer to set identification as partial identification.

model, this is not the case in higher-dimensional partially identified models. In general, all we are imposing is that the sign pattern of each of the unidentified shocks is different from that of the identified shocks. We are not imposing that any of the unidentified shocks must have a specific sign pattern.

It should be noted that there is no consensus in the literature on whether this additional requirement should be imposed in estimating partially identified sign-identified models. As a result, some studies focus only on the responses to the shock of interest without inspecting the responses to other shocks in the same model. In other words, they consider only one column of the impact multiplier matrix at a time, looking for solutions that satisfy the prespecified sign pattern while ignoring the other columns. All columns satisfying the prespecified sign pattern are considered admissible solutions, and the possibility that other shocks in the same structural model may have the same sign pattern is ignored. Further discussion of this issue can be found in Uhlig (2005), Fry and Pagan (2011), and Canova and Paustian (2011).

13.4 Beyond Static Sign Restrictions

The basic algorithm for characterizing the set of admissible models can be extended to allow for additional restrictions on the structural impulse responses. Such additional restrictions are often required for the estimates of sign-identified models to be economically meaningful.

13.4.1 Dynamic Sign Restrictions

It is straightforward to extend the algorithms above to allow for additional sign restrictions on the structural impulse responses beyond the impact period. One drawback of this approach is that there is less consensus in economic theory about the signs of structural impulse responses at longer horizons (see, e.g., Canova and Paustian 2011). It has also been noted that imposing dynamic sign restrictions may be redundant in some cases and hence unhelpful (see Fry and Pagan 2011). Nevertheless, such restrictions can be useful in restricting further the space of admissible models. For example, we may be willing to agree that a monetary policy tightening is bound to reduce real GDP after half a year, even if the sign of the short-run response and the sign of the longer-run response of real GDP is debatable (see Inoue and Kilian 2013).

13.4.2 Elasticity Bounds

A common approach among applied researchers has been to favor models that are agnostic in the sense of only imposing minimal identifying assumptions.

Being agnostic in the context of sign-identified models involves the risk of allowing problematic structural models to be retained in the admissible set. This point was first illustrated in Kilian and Murphy's (2012) analysis of the global oil market. Their model may be viewed as the analogue of the recursively identified oil market model of Kilian (2009) in Chapter 8.

The model includes monthly data for the growth of global crude oil production ($\Delta prod_t$), a measure of global real activity expressed as a business cycle index (rea_t), and the log of the real price of oil ($rpoil_t$). The three structural shocks are a shock to the flow of crude oil coming out of the ground ($w_t^{\text{flow supply}}$), a shock to the flow demand for oil associated with unexpected fluctuations in the global business cycle ($w_t^{\text{flow demand}}$), and a shock to the demand for oil not associated with the global business cycle ($w_t^{\text{other demand}}$). The latter demand shock is designed to capture, for example, shifts in the precautionary demand for crude oil that shift the demand for stocks of oil. A conventional analysis of this market would start with restrictions on the signs of the structural impact multiplier matrix:

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} - & + & + \\ - & + & - \\ + & + & + \end{bmatrix} \begin{pmatrix} w_t^{\text{flow supply}} \\ w_t^{\text{flow demand}} \\ w_t^{\text{other demand}} \end{pmatrix}.$$

All shocks have been normalized to imply an increase in the real price of oil on impact. For example, a negative flow supply shock can be thought of as a shift of the short-run oil supply curve to the left, causing a reduction in oil production and an increase in the real price of oil. Such a shock would also lower global real economic activity. In contrast, a positive flow demand shock would shift the demand curve for oil to the right, causing both oil production and the real price of oil to increase while raising global real activity. Finally, we can think of the other demand shock as a precautionary demand shock that raises the price of oil by shifting the demand for oil stocks. To accommodate the accumulation of stocks, oil production has to increase and oil consumption has to fall, which implies a fall in global real activity.

One of the central questions in the literature on the global oil market is how large the effects of flow supply shocks are on the real price of oil. It can be shown that the set of admissible models consistent with the minimal assumptions outlined thus far includes many models in which oil supply shocks cause much larger oil price responses than in the recursively identified model of Kilian (2009).

What is not immediately obvious, however, is that these models all imply a large impact price elasticity of oil supply. The latter elasticity is defined as the ratio of the impact response of global oil production triggered by an exogenous demand shock relative to the impact response of the real price of oil triggered

by the same shock.[3] Large elasticity estimates not only violate the conventional wisdom among oil experts but also are inconsistent with formal and informal empirical evidence from other sources showing that this elasticity is close to zero. They are also inconsistent with economic theory (see, e.g., Anderson, Kellogg, and Salant 2016).

Kilian and Murphy (2012) show that restricting the impact price elasticity of oil supply not to exceed a reasonable magnitude suffices to eliminate the structural models in the admissible set that imply large responses of the real price of oil to flow supply shocks and produces results much more in line with those in Kilian (2009). Although this additional restriction does not directly restrict the response of the real price of oil to a supply shock, it suffices to estimate fairly precisely the set of responses of the real price of oil to oil supply shocks. In contrast, there is considerable uncertainty about the magnitude and pattern of the responses of the real price of oil to the two demand shocks in this model.

This example shows how a failure to impose relevant identifying restrictions may give credence to models that upon reflection are unrealistic. Kilian and Murphy (2012) conclude that the burden of proof is on the researcher to show that all available identifying information has been imposed on sign-identified models. There is no comfort in having remained agnostic and no excuse for having failed to examine other possible sources of identifying information.

In practice, elasticity bounds may be imposed by disregarding the candidate solutions for all structural models with higher elasticity values than permitted. The approach of bounding price elasticities was subsequently refined in Kilian and Murphy (2014) who showed that the impact price elasticity of oil demand as well may be bounded by independent estimates of the long-run price elasticity of demand from cross-sectional studies. The identifying assumption is that the long-run price elasticity of oil demand at least weakly exceeds the short-run price elasticity. Although this example focuses on the oil market, similar restrictions may also be used in many other market models. Even if a specific bound may be difficult to justify, sensitivity analysis based on alternative elasticity bounds may reveal how fragile the empirical results are to imposing extra information.

Although the example of market models is the most natural context for the use of elasticity bounds, such bounds have also been discussed in other contexts. For example, Caldara and Kamps (2017) investigate the implications of bounds on output elasticities of fiscal variables for the fiscal multiplier in sign-identified VAR models.

[3] This definition corresponds to the textbook definition of the price elasticity of oil supply. In applied regression analysis it is common to regress the log of one variable on the log of another variable and to refer to the regression coefficient as an elasticity. The latter elasticity concept in general is not related to the microeconomic concept of the price elasticity of demand or the price elasticity of supply.

13.4.3 Shape Restrictions

A natural extension of the idea of sign restrictions is the use of shape restrictions on structural impulse response functions. Such restrictions may involve monotonicity constraints on the evolution of the response function or may involve imposing a hump shape on the response function. A hump shape may be natural in modeling the effects of demand shocks on real output, for example (see, e.g., Blanchard and Quah 1989; Inoue and Kilian 2016). Shape restrictions may also arise in modeling the delayed overshooting in asset prices in response to monetary policy shocks, as discussed in Eichenbaum and Evans (2005) and Scholl and Uhlig (2008). These restrictions can be written as sign restrictions on the change in the impulse response coefficients across the horizon and hence can be accommodated by standard solution algorithms.

13.5 Can Sign Restrictions Be Verified?

It is uncontroversial that imposing incorrect identifying assumptions could result in the set of admissible models being empty. Perhaps motivated by this observation, it is sometimes suggested that only a small fraction of the random draws of PQ being admissible (in the sense of satisfying the maintained sign restrictions) means that the identifying assumptions are suspect. For example, Fry and Pagan (2011) suggest that in this case the data are incompatible with the maintained sign restrictions.

This view is misleading for several reasons. First, suppose for expository purposes that the reduced-form parameters are known. Then the fraction of admissible models only depends on Q, but Q does not depend on the data, so the fraction of admissible models does not constitute empirical evidence. In fact, all rotations by construction imply models that fit the data equally well. Second, reporting the fraction of admissible models is meaningless, if we do not know the computational efficiency of the algorithm used. Merely by changing algorithms, this fraction may change by a factor of 10 or 20. Third, it is evident upon reflection that, for a given algorithm, the fraction of admissible models is simply an indication of how informative the identifying restrictions are. This last point is best made by example.

To build intuition, first consider an extreme example, in which the data are generated from a recursive model. In that case, no one would expect random draws from the QR algorithm conditional on the VAR parameters to generate realizations of PQ that preserve this recursive structure. Although the probability of encountering a recursive model estimate is zero, the correct model is nevertheless recursive. The same intuition still applies if we relax the assumptions somewhat. Consider instead a set of data generating processes with some structural impulse responses not equal to zero, but ranging in value between 0 and some small $\epsilon > 0$ in population. If we impose this constraint in our search for

admissible models, the number of draws generated from the QR decomposition that lie within this region will be small relative to the total number of draws by construction. Again, this does not mean that the identifying assumption is suspect. Exactly the same phenomenon arises if – rather than restricting some structural impulse response to lie within a small region – one imposes multiple inequality restrictions on the impulse responses that effectively restrict the admissible range of the structural impulse responses. Indeed, a low fraction of admissible models in this case could actually be an indication of having imposed all relevant economic structure on the model. Conversely, a high fraction of admissible models is bound to arise in insufficiently identified models. Thus, the fraction of admissible structural models is not a meaningful measure of the quality of the identifying restrictions of a sign-identified VAR model.

It is, of course, correct that by imposing a certain sign on the structural impulse responses, we rule out all models that imply a different sign. As in all structural VAR models, the estimates of sign-identified models are conditional on the chosen identifying assumptions. These assumptions are not testable within this VAR framework. The only purpose of estimating the sign-identified model therefore is to quantify the magnitude of the response of interest conditional on the assumed signs.

It has been suggested that one way of evaluating the plausibility of sign restrictions would be to establish that the sign-identified model can recover the correct sign of the response of interest, when other response functions in the VAR model are constrained, but not the response function of interest. Indeed, this was the explicit motivation for the agnostic identification procedure proposed by Uhlig (2005). Paustian (2007) addresses this question in the context of sign-identified VAR models applied to data generated from DSGE models. He provides simulation evidence that for this approach to work, the variance of the structural shock that triggers the response of interest must be much higher in population than seems empirically plausible.

This result is not surprising because sign restrictions are comparatively weak, and hence models that do not impose all available identifying information tend to be uninformative. This point has been illustrated by Inoue and Kilian (2013) in the context of Uhlig's agnostic identification procedure. Their analysis confirms that there is no information about the response of real GDP to monetary policy shocks in this model without further identifying restrictions on the response function of interest. This result, of course, does not mean that sign-identified models with richer identification structures are not a useful tool in quantifying dynamic economic relationships. A similar point has been made in Canova and Paustian (2011) who conclude that – leaving aside pathological examples – even when the variance of the shock of interest is low, the sign-identified model usually does not make systematic mistakes, provided the estimated model has a sufficiently rich shock structure and employs enough identifying restrictions.

13.6 Estimation and Inference in Sign-Identified VAR Models

A fundamental problem in interpreting VAR models identified based on sign restrictions is that there is not a unique point estimate of the structural impulse response functions. Unlike conventional structural VAR models based on short-run restrictions, sign-identified VAR models are only set identified. This problem arises because sign restrictions represent inequality restrictions. The cost of remaining agnostic about the precise values of the structural model parameters is that the data are potentially consistent with a wide range of structural models that are all admissible in that they satisfy the identifying restrictions. Without further assumptions all these structural models are equally likely.

A typical outcome in practice is that the structural impulse responses implied by the admissible structural models will disagree on the substantive economic questions of interest. One early approach to this problem, exemplified by Faust (1998), has been to focus on the admissible model that is most favorable to the hypothesis of interest. This allows us to establish the extent to which this hypothesis could potentially explain the data. It may also help us rule out a hypothesized explanation if none of the admissible models supports this hypothesis. The problem is that this approach is not informative about whether any one of the admissible models is a more likely explanation of the data than some other model. There are examples in which the admissible structural models are sufficiently similar to allow unambiguous answers to the question of economic interest (see, e.g., Kilian and Murphy 2012). Typically, however, the set of admissible models will be equally consistent with competing economic hypotheses. This problem is compounded once we allow for estimation uncertainty in the reduced-form parameters. Moreover, in the latter case, it has to be kept in mind that confidence sets or credible sets in sign-identified VAR models by construction reflect both uncertainty about the identification of the model and estimation uncertainty.

There have been both frequentist and Bayesian approaches to summarizing estimates of the admissible set of sign-identified structural VAR models. These approaches differ from the methods discussed in Chapter 12. Standard asymptotic and bootstrap methods of inference on structural impulse responses, in particular, are invalid when working with inequality restrictions. This result follows from the absence of a consistent point estimator of the structural impulse response. Section 13.6.1 reviews the construction of alternative frequentist confidence sets for structural impulse responses in sign-identified VAR models. Section 13.6.2 discusses how to generate draws from the posterior distribution of the structural impulse responses. Section 13.6.3 critically examines the proposal of summarizing the posterior with the help of so-called median response functions and discusses the alternative approach of reporting the responses of the most likely structural model in the identified set. It also

discusses the construction of credible sets. Yet another approach is the use of penalty functions, as reviewed in Section 13.6.4. Finally, Section 13.6.5 outlines how historical information may be used to narrow the set of admissible models both in a Bayesian and in a frequentist setting.

13.6.1 Frequentist Approaches

The procedures for characterizing the set of admissible models outlined in Sections 13.2 and 13.4 condition on a given reduced-form VAR model and do not account for estimation uncertainty. In practice, we need to account for both identification uncertainty and estimation uncertainty in sign-identified models. One method of constructing classical confidence intervals for structural impulse responses in stationary sign-identified VAR models with iid innovations has recently been developed by Moon, Schorfheide, and Granziera (2013) under the simplifying assumption that the reduced-form impulse responses have an asymptotic normal distribution. They view the estimation of sign-identified VAR models as an estimation problem subject to moment-inequality restrictions. The presence of inequality constraints means that the structural impulse responses are only set identified and that the large-sample numerical equivalence between Bayesian credible sets and frequentist confidence sets breaks down.

Let ϕ denote the vector of orthogonalized reduced-form impulse responses, obtained by post-multiplying the reduced-form VAR responses by the lower-triangular Cholesky decomposition of Σ_u, θ the vector of structural impulse responses, and q the vector of nuisance parameters corresponding to a suitable parameterization of Q. Sign restrictions generate an identified set for θ and q, denoted by $F^{\theta,q}(\phi)$. Conditional on q and ϕ, the vector θ is point identified. Let $\sqrt{T}(\widehat{\phi} - \phi) \xrightarrow{d} \mathcal{N}(0, \Lambda)$ for all $\phi \in \mathcal{P}$, where $\mathcal{P}$ is the space of the reduced-form parameters ϕ and the limiting covariance matrix Λ is of full rank.

Moon et al. propose two approaches to obtaining a $(1-\gamma)100\%$ confidence set CS^{θ}

$$\inf_{\phi \in \mathcal{P}} \inf_{\theta \in F^{\theta}(\phi)} \mathbb{P}_{\phi} \left\{ \theta | \theta \in \text{CS}^{\theta}(\widehat{\phi}) \right\} \geq 1 - \gamma,$$

where $\mathbb{P}_{\phi}$ denotes the probability measure for a given reduced-form draw of ϕ and $F^{\theta}(\phi)$ is the set of structural impulse responses obtained for a given ϕ. Both approaches involve the marginalization of the joint confidence set for (θ, q). One approach is to project the joint confidence set for $(\theta, q) \in F^{\theta,q}(\phi)$ onto the θ ordinate. The other approach involves first constructing confidence intervals for the set-identified nuisance parameters in q and then taking the union of standard confidence sets for θ that are generated conditional on all q in the first-stage confidence set. Moon et al. use the Bonferroni inequality to control the coverage probability of the resulting confidence set for θ.

In practice, the distribution of the reduced-form parameters is approximated by standard bootstrap methods. A detailed description of the algorithms is provided in Moon, Schorfheide, and Granziera (2013). The construction of these nonstandard confidence intervals is computationally costly and tends to be infeasible in fully identified models. Moreover, their coverage probabilities tend to be conservative.

More recently, Gafarov, Meier, and Montiel Olea (2015a) proposed a computationally less demanding alternative delta method approach for partially identified set-identified structural VAR models that tends to produce tighter intervals than the algorithm of Moon et al. A more general method that also accommodates fully identified models is discussed in Gafarov, Meier, and Montiel Olea (2015b). The latter method is conservative in that its coverage tends to exceed the nominal confidence level, but the coverage accuracy may be calibrated, resulting in tighter confidence bounds. Unlike the method in Moon, Schorfheide, and Granziera (2013), Gafarov, Meier, and Montiel Olea (2015a, 2015b) only require the asymptotic normality of the reduced-form parameters and are able to achieve uniformly accurate coverage.

Even with these adjustments, frequentist confidence sets for sign-identified models tend to be wide and not informative about the shape of the impulse response functions. A given confidence band tends to be consistent with a wide range of different impulse response function patterns. This fact tends to make it difficult to interpret the results from an economic point of view. An open question is whether this problem may be overcome in larger models with more identifying restrictions.

Finally, Kitagawa, Montiel Olea, and Payne (2015) extend the analysis to the related problem of constructing confidence sets for the maximum of the h-period ahead forecast error variance decomposition of a variable with respect to a given structural shock (see Chapter 4). The key innovation of this study is that it focuses on possibly nondifferentiable functions of a parameter vector $\boldsymbol{\theta}$ such as $|\boldsymbol{\theta}|$ or $\max(0, \boldsymbol{\theta})$, for which standard delta method or bootstrap inference, as employed in earlier studies, fails. The maximal contribution of monetary policy shocks to the variability of output growth is one such example. Kitagawa et al. provide precise conditions under which a confidence set for such functions may nevertheless be obtained by calibrating the highest-posterior density Bayesian credible set to achieve the desired frequentist coverage accuracy.

Notwithstanding these theoretical advances, there are few empirical applications to date of frequentist confidence sets for sign-identified structural impulse responses. Moreover, it is important to keep in mind that these confidence sets not only tend to be wide, but that they tend to be uninformative about the shape of the impulse response functions in that a given confidence band is typically consistent with a wide range of different impulse response function patterns. This fact tends to make it difficult to interpret the results from an economic point of view.

13.6.2 Bayesian Approaches

The most common approach in the literature on sign-identified VAR models has been to rely on Bayesian methods of inference (see Chapter 5). For example, under the assumption of a conventional Gaussian-inverse Wishart prior on the reduced-form parameters and an independent uniform prior on the rotation matrices, one can construct the posterior distribution of the structural impulse responses by simulating posterior draws from the reduced-form posterior, applying the QR algorithm to each reduced-form posterior draw, and discarding solutions that do not satisfy the sign restrictions.

Generating the Posterior in Fully Identified Models. Consider the K-variate reduced-form VAR(p) model:

$$y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + u_t, \tag{13.6.1}$$

for $t = 1, \ldots, T$, where $u_t \overset{iid}{\sim} \mathcal{N}(0_{K\times 1}, \Sigma_u)$ and Σ_u is positive definite. Define $A = [\nu, A_1, \ldots, A_p]$.

Specify a Gaussian-inverse Wishart prior distribution for the reduced-form VAR parameters of the form

$$\text{vec}(A)|\Sigma_u \sim \mathcal{N}(\text{vec}(A^*), V_{\text{vec}(A)} = V \otimes \Sigma_u)$$

and

$$\Sigma_u \sim \mathcal{IW}_K(S_*, n),$$

where $\text{vec}(A^*)$, V, S_*, and n are prior parameters specified by the analyst (see Section 5.2.4). Then the posterior is given by

$$\text{vec}(A)|\Sigma_u, \mathbf{y} \sim \mathcal{N}(\text{vec}(\bar{A}), \bar{\Sigma}_{\text{vec}(A)}), \quad \Sigma_u|\mathbf{y} \sim \mathcal{IW}_K(S, \tau), \tag{13.6.2}$$

where $\mathbf{y} \equiv \text{vec}(Y)$, $Y \equiv [y_1, \ldots, y_T]$ denotes the data,

$$\begin{aligned}
\bar{A} &= (A^* V^{-1} + YZ')(V^{-1} + ZZ')^{-1} \\
\bar{\Sigma}_{\text{vec}(A)} &= (V^{-1} + ZZ')^{-1} \otimes \Sigma_u, \\
S &= T\widetilde{\Sigma}_u + S_* + \hat{A}ZZ'\hat{A}' + A^* V^{-1} A^{*\prime} - \bar{A}(V^{-1} + ZZ')\bar{A}',
\end{aligned}$$

and $\tau = T + n$. Here A^* and $\bar{A}$ are $K \times (Kp + 1)$ matrices, $\hat{A} = YZ'(ZZ')^{-1}$, and $\widetilde{\Sigma}_u = (Y - \hat{A}Z)(Y - \hat{A}Z)'/T$. Moreover, $Z \equiv [Z_0, \ldots, Z_{T-1}]$ with $Z_{t-1} \equiv (1, y'_{t-1}, \ldots, y'_{t-p})'$.

Simulating the posterior of the structural impulse responses requires draws for $\text{vec}(A)$ and for B_0^{-1}. Let A^{*r} denote the r^{th} posterior draw of A and Σ_u^{*r} the r^{th} posterior draw for Σ_u. Then $\widetilde{B}_0^{-1} = P^{*r}Q$, where P^{*r} is the lower-triangular Cholesky decomposition of Σ_u^{*r} such that $P^{*r}P^{*r\prime} = \Sigma_u^{*r}$, and Q is an orthogonal matrix. $\widetilde{B}_0^{-1}$ is a potential solution for the unknown structural impact multiplier matrix B_0^{-1} that satisfies $\widetilde{B}_0^{-1}\widetilde{B}_0^{-1\prime} = P^{*r}QQ'P^{*r\prime} = P^{*r}P^{*r\prime} = \Sigma_u^{*r}$.

Following Uhlig (2005), the prior distribution for the matrix Q is postulated to be uniform on the space of orthogonal matrices $\mathcal{O}(K)$, which can be drawn via the Householder transformation as described in Section 13.2.2, allowing us to simulate the set of potential solutions for $\widetilde{B}_0^{-1}$, given Σ_u^{*r} and A^{*r}.

As discussed in Chapter 4, there is a known nonlinear function that allows us to construct the structural impulse responses associated with every potential solution $(A^{*r}, \widetilde{B}_0^{-1})$ for the structural model. Candidate solutions that imply structural impulse responses that do not satisfy all sign restrictions are discarded. The other posterior draws for the structural impulse responses are retained and used to approximate the posterior distribution of the structural impulse responses. Unlike in conventional VAR models, this distribution reflects both estimation uncertainty and identification uncertainty.

In practice we proceed as follows:

Step 1. Take a random draw, (A^{*r}, Σ_u^{*r}), from the posterior of the reduced-form VAR parameters and compute the lower-triangular Cholesky decomposition $P^{*r} = \text{chol}(\Sigma_u^{*r})$.

Step 2. For (A^{*r}, P^{*r}), consider N random draws of the rotation matrix Q, and for each combination (A^{*r}, P^{*r}, Q) compute the set of implied structural impulse responses Θ^{*r}.

Step 3. If Θ^{*r} satisfies the sign restrictions, store the value of Θ^{*r}. Otherwise discard Θ^{*r}.

Step 4. Repeat steps 1, 2, and 3 M times.

In simulating the posterior distribution of the structural responses, care must be taken that the posterior is approximated using a sufficiently large number of reduced-form draws as well as a sufficiently large number of rotation draws for each posterior draw from the reduced-form. The tighter the identifying restrictions, the more draws will be required to approximate the posterior of the structural impulse responses, because fewer realizations will satisfy the sign restrictions. The precise number of draws required for an accurate approximation also depends on which algorithm is employed. In practice, it is recommended to verify that alternative seeds of the Gaussian random number generator generate similar sets of admissible models. If not, the number of draws must be increased.

Generating the Posterior in Partially Identified Models. A simplified algorithm for partially identified models was proposed by Uhlig (2005). Uhlig's objective was to identify the responses to a monetary policy shock without taking a stand on the identification of the remaining structural shocks. Uhlig does not solve his model by constructing draws for Q as discussed earlier. Rather, he observes that constructing the responses of the model variables to the monetary policy shock does not require knowledge of the entire structural

impact multiplier matrix B_0^{-1}, but only of the vector b representing the impact responses of his model variables to the monetary policy shock. The vector b is the first column of B_0^{-1} in this example.

Focusing on one structural shock only simplifies the analysis. By analogy to $B_0^{-1} = PQ$, let q denote a K-dimensional vector of unit length such that $b = Pq$. Then the sign-identified structural impulse responses can be expressed as $\theta_i = \Phi_i b$, where θ_i denotes the vector of responses at horizon i to the monetary policy shock. In practice, Uhlig draws a vector $\widetilde{q}$ from the $\mathcal{N}(0, I_K)$ distribution and normalizes the length of this vector to unity such that $q = \widetilde{q}/\|\widetilde{q}\|$, where $\|\widetilde{q}\| = \sqrt{\sum_{k=1}^{K} \widetilde{q}_k^2}$ denotes the length of the vector $\widetilde{q}$. Given P and a candidate solution for q, one can construct the implied candidate solution for b. Only candidates for b that imply solutions for θ_i that satisfy the sign restrictions are retained.

This approach does not rule out the possibility that there are other orthogonal shocks with exactly the same sign pattern as the monetary policy shock, which may raise the question of whether the monetary policy shock is properly identified.

13.6.3 Evaluating the Posterior of the Structural Impulse Responses

Having obtained enough draws for the structural responses to approximate their posterior distribution, we can make probability statements about the structural impulse responses. Note that the posterior distribution incorporates both estimation uncertainty and identification uncertainty. The standard approach in the literature for many years has been to report the vector of pointwise posterior medians of the structural impulse responses (often referred to as the median response function) as a measure of the central tendency of the impulse response functions. In fact, many applied users treat these median response functions as though they were traditional point estimates from exactly identified models.

This approach suffers from two distinct shortcomings. One shortcoming is that the vector of pointwise posterior median responses will not correspond to the response function of any of the admissible models, unless the pointwise posterior medians of all impulse response coefficients in the VAR system correspond to the same structural model, which is highly unlikely a priori. The problem is not only that for different horizons h the pointwise posterior median responses coincide with responses from different admissible structural models. Similar problems may also arise when the order of the models differs for two response functions at the same horizon h. Thus, the median response function lacks a structural economic interpretation (see, e.g., Fry and Pagan 2011; Kilian and Murphy 2012; Inoue and Kilian 2013).

This point is illustrated in Figure 13.6. The figure focuses on the response of real GDP to an unanticipated monetary policy shock for a horizon of up

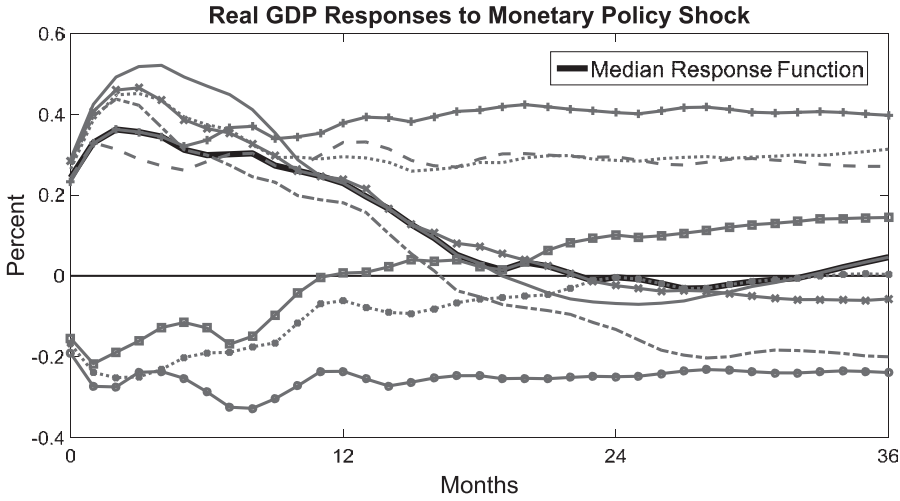

Figure 13.6. Randomly selected response functions from the sign-identified VAR model in Uhlig (2005).

to 36 months. This example was constructed by plotting a randomly chosen subset of nine admissible response functions for the sign-identified VAR(12) model of monetary policy proposed by Uhlig (2005). The model consists of monthly U.S. data for the log of interpolated real GDP, the log of the interpolated GDP deflator, the log of a commodity price index, total reserves, nonborrowed reserves, and the federal funds rate. The sample period is 1965m1-2003m12. The identifying restrictions are that an unexpected monetary policy contraction is associated with an increase in the federal funds rate, with price decreases, and with declines in nonborrowed reserves for some time following the policy shock.

Figure 13.6 demonstrates that, for different horizons, the pointwise posterior median responses of real GDP to a monetary policy shock coincide with responses of different admissible structural models. Specifically, the median response function for real GDP coincides with the response function of model 1 at horizons 23-32, with the response function of model 3 at horizons 19 and 20, with that of model 5 at horizons 12-18 and 33-36, with that of model 6 at horizons 0, 1, and 7-9, with that of model 7 at horizons 10, 11, 21, and 22, with that of model 8 at horizons 5 and 6, and with that of model 9 at horizons 2, 3, and 4. There is, in fact, no structural model in the admissible set that could replicate the response pattern implied by the posterior median response function, rendering this statistic economically meaningless.

The second shortcoming is that median response functions are not a useful statistical summary of the set of admissible impulse response functions. It is well known that the vector of medians is not the median of a vector valued random variable. In fact, the median of a vector valued random variable does not

exist, rendering the vector of pointwise medians inappropriate as a statistical measure of the central tendency of the impulse response functions (see, e.g., Chaudhuri 1996; Koltchinskii 1997; Liu, Parelius, and Singh 1999).[4]

This means that even if there were an admissible structural model for which all impulse response functions coincide with the corresponding median response functions, there would be no compelling reason to focus on this model in interpreting the evidence. In fact, it has been shown that posterior median response functions may be quite misleading about the most likely response dynamics in sign-identified models (see, e.g., Kilian and Murphy 2012; Inoue and Kilian 2013).

This criticism also extends to the proposal in Fry and Pagan (2011) designed to overcome the lack of structural interpretation of median response functions. Their proposal is to search for the admissible structural model with impulse response functions closest to the median response function. This proposal deals with the first shortcoming highlighted above, but not with the second. Because the median response function is not a well-defined statistical measure of central tendency, there is no compelling reason to focus on the structural model with responses closest to the pointwise posterior medians.

The same conceptual problem arises with the upper and lower quantiles constructed from the pointwise posterior distribution of the structural impulse responses. It is common to connect the upper quantiles for a given impulse response function to form an upper band, and similarly to connect the lower quantiles to form a lower band. These error bands, of course, fail to account for the dependence of structural impulse responses across horizons and across response functions, and hence are misleading (see, e.g., Sims and Zha 1999).

A solution to these two problems within the Bayesian framework of Uhlig (2005) has been proposed in Inoue and Kilian (2013). This study shows how to characterize the most likely admissible model within the set of structural VAR models that satisfy the sign restrictions. The most likely structural model can be computed from the posterior mode of the joint distribution of admissible models both in the fully identified and in the partially identified case. The resulting set of structural response functions is well defined from an economic and a statistical point of view. Inoue and Kilian also propose a highest posterior density credible set that characterizes the joint uncertainty about the set of admissible models. Unlike conventional posterior error bands for sign-identified VAR models, the implied credible sets for the structural response functions characterize the full uncertainty about the structural response functions.

[4] It is worth noting that this problem persists even if we restrict attention to the structural responses at horizon 0 because $\text{vec}(B_0^{-1})$ is a multdimensional object. Put differently, the posterior median for one element of $\text{vec}(B_0^{-1})$ may come from a different structural model than the posterior median of another element of $\text{vec}(B_0^{-1})$.

The procedure proposed by Inoue and Kilian (2013, 2017) can be summarized as follows. We first consider the case of a fully identified model, before extending the discussion to partially identified models.

Fully Identified Models. The objective is to rank the structural models in the admissible set. Ignoring the intercept, which does not enter the definition of the structural impulse responses, let $A \equiv [A_1, \ldots, A_p]$, let P be the lower-triangular Cholesky decomposition of Σ_u and let vech(P) denote the $K(K+1)/2 \times 1$ vector that consists of the on-diagonal elements and the below-diagonal elements of P. The $K \times K$ real orthogonal matrix $Q \in \mathcal{O}(K)$ satisfies $Q'Q = I_K$ and its determinant is $|Q| = 1$. Then

$$\mathbf{S} = I_K - 2(I_K + Q)^{-1}$$

is a $K \times K$ skew-symmetric matrix. Let $\mathbf{s}$ denote the $K(K-1)/2 \times 1$ vector that consists of the above-diagonal elements of $\mathbf{S}$. Then the matrix Q and the vector $\mathbf{s}$ have a one-to-one relationship. If Q is uniformly distributed over the space of real orthogonal $K \times K$ matrices $\mathcal{O}(K)$, the density of $\mathbf{s}$ is given by

$$f(\mathbf{s}) = \left(\prod_{i=2}^{K} \frac{\Gamma(i/2)}{\pi^{i/2}} \right) \frac{2^{(K-1)(K-2)/2}}{|I_K + \mathbf{S}|^{K-1}},$$

where $\Gamma(x) \equiv \int_0^\infty z^{x-1} e^{-z} dz$ and $|.|$ denotes the determinant (see equation (4) of León, Massé, and Rivest 2006, p. 415).

Let $B_0^{-1} = PQ$. Then there is a one-to-one mapping between the first $p+1$ structural impulse responses

$$\Theta = [B_0^{-1}, \Phi_1 B_0^{-1}, \Phi_2 B_0^{-1}, \ldots, \Phi_p B_0^{-1}],$$

on the one hand, and the tuple

$$(\text{vec}(A), \text{vech}(P), \mathbf{s})$$

on the other hand. This one-to-one mapping allows us to derive the posterior density $f(\Theta)$ from the joint posterior density of (vec(A), vech(P), $\mathbf{s}$), where Θ is defined by the known nonlinear function $\Theta = h(\text{vec}(A), \text{vech}(P), \mathbf{s})$. Using the change-of-variables method, the posterior density of Θ can be written in closed form as

$$f(\Theta) \propto \left| \frac{\partial \text{vec}(\Theta)}{\partial [\text{vec}(A)', \text{vech}(P)', \mathbf{s}']} \right|^{-1} \left| \frac{\partial \text{vech}(\Sigma_u)}{\partial \text{vech}(P)'} \right| f(A|\Sigma_u) f(\Sigma_u) f(\mathbf{s}).$$

This result allows one to compute the numerical value of the posterior density for every possible draw Θ of the structural model. Let $\boldsymbol{\Theta}$ denote the set of all admissible structural models that satisfy the sign restrictions. If the objective is to choose one admissible structural model from all possible realizations in $\boldsymbol{\Theta}$, then a natural approach is to select the modal (or most likely) model, defined as the structural model that maximizes the value of $f(\Theta)$ among all

models that satisfy the sign restrictions. By construction this model is immune from both shortcomings of the posterior median response functions discussed earlier. It must be emphasized that Inoue and Kilian (2013) do not propose to focus on the mode of the marginal distribution of the structural impulse responses, but rather on the mode of the joint distribution of the structural models.

The corresponding $(1-\gamma)100\%$ highest posterior density (HPD) credible set may be defined by

$$\mathcal{S} = \{\Theta \in \mathbf{\Theta} | f(\Theta) \geq c_\gamma\}, \tag{13.6.3}$$

where $f(\Theta)$ is the posterior density of Θ and c_γ is the largest constant such that

$$\mathbb{P}(\mathcal{S}) \geq 1-\gamma.$$

This credible set is a joint credible set accounting for the dependence of the elements of Θ across time as well as across variables.

In practice we proceed as follows:

Step 1. Take a random draw, (A^{*r}, P^{*r}), from the posterior of the reduced-form VAR parameters.

Step 2. For (A^{*r}, P^{*r}), consider N random draws of the rotation matrix Q and for each combination (A^{*r}, P^{*r}, Q) compute the set of implied structural impulse responses Θ^{*r}.

Step 3. If Θ^{*r} satisfies the sign restrictions, store the value of Θ^{*r} and the value of $f(\Theta^{*r})$. Otherwise discard Θ^{*r}.

Step 4. Repeat steps 1, 2, and 3 M times. Sort in descending order by the value of $f(\Theta^{*r})$ the pairs $\{(\Theta^{*r}, f(\Theta^{*r}))\}$ that satisfy the sign restrictions. Then the Θ^{*r} in the first sorted pair is the most likely model and the $(1-\gamma)100\%$ HPD credible set is obtained by selecting the Θ^{*r} of the first $(1-\gamma)100\%$ sorted pairs.

Inoue and Kilian (2013) make the case for focusing on the mode of the posterior distribution of the admissible structural models (with each model characterized by an entire set of structural impulse responses) as opposed to the pointwise median of the structural impulse response vector. It may seem that instead of focusing on the posterior mode in the admissible set of models one could have reported the posterior mean of the responses of the admissible structural models (or the responses of the admissible structural model whose impulse responses are closest to that posterior mean), given that the mean is a well-defined measure of central tendency for vector-valued random variables. There are two reasons for not proceeding along these lines. One is that it is not clear how to conduct inference about the posterior mean given that $f(\Theta)$ is highly non-Gaussian. The other is that the distribution in question need not have finite moments, as discussed in Baumeister and Hamilton (2015a).

The case has been made that the median response function may be motivated as the solution to a loss function involving the sum of the absolute loss of the vector of structural responses (see, e.g., Baumeister and Hamilton 2015a). While the median response function is indeed optimal conditional on this particular loss function, there are many possible loss functions one could maintain. The real question therefore is not whether there is a loss function under which the median response function is optimal, but rather what the right loss function is from the point of view of an economist evaluating the VAR model. It is not enough for a summary statistic to be statistically coherent. It also has to be economically meaningful. The loss function proposed by Baumeister and Hamilton (2015a), under which the median response function is optimal, effectively postulates that the economist using this structural VAR model is not interested in the shapes of the impulse response functions or in their comovement. Reporting pointwise posterior quantiles thus does not account for the dependence of structural impulse responses over time and across variables. Hence, the case can be made that Inoue and Kilian's loss function is more in line with the objectives of applied users. Moreover, the questions answered by their methodology, namely what the most likely structural models are within the admissible set, cannot be answered by computing the median response function.

Partially Identified Models. If we are concerned with a subset of structural impulse response functions only, as in the Uhlig (2005) example, what matters for constructing the posterior mode is not the joint impulse response distribution, but the marginalized distribution obtained by integrating out responses to shocks that are not economically identified. To simplify the exposition, we focus on the case in which only impulse responses to one structural shock are identified. Let b denote the column vector of B_0^{-1} relating to this shock and denote by q a random vector drawn from $\mathcal{N}(0, I_K)$ and normalized to have unit length. In other words, $\sum_{i=1}^{K} q_i^2 = 1$, where q_i refers to element i of the vector q. The K-dimensional vector b represents the impact responses of the model variables to the shock in question. It is computed as $b = Pq$ by analogy to $B_0^{-1} = PQ$. The implied structural impulse responses at higher horizons are $\theta_1 = \Phi_1 b, \ldots, \theta_p = \Phi_p b$. Given $\theta_0 = b$ and Kp restrictions of the form

$$\theta_i = \Phi_i b, \quad i = 1, \ldots, p,$$

the derivation of the marginal posterior density $f(\theta_0, \theta_1, \ldots, \theta_p)$ requires numerical integration (see Inoue and Kilian 2013). On the basis of this density one may proceed exactly as in the fully identified case. In practice, this density must be evaluated by Monte Carlo integration, which greatly increases the computational cost compared with the algorithm for fully identified models.

Of course, as noted before, this approach does not ensure that the remaining structural shocks have a sign pattern that is distinct from the shock of interest.

13.6.4 The Penalty Function Approach

An alternative approach to selecting one model from the set of admissible structural models is to minimize a criterion function (or penalty function). This approach was first discussed in Uhlig (2005). Uhlig proposed to select from the set of candidate models the structural model that minimizes a penalty function. Rather than directly discarding models that violate sign restrictions, as in the conventional sign restriction approach, Uhlig attaches a large numerical penalty to such models. His penalty function not only punishes for violations of sign restrictions more than it rewards models with correct signs, but it involves an additional reward for large responses which may be tailored to specific responses and horizons. For example, a model in which monetary policy shocks generate a large initial increase in the federal funds rate, all else equal, may receive a larger reward than other models.

Uhlig's identifying restrictions are that an unexpected monetary policy contraction is associated with an increase in the federal funds rate, with price decreases, and with declines in nonborrowed reserves for some time following the policy shock. Uhlig's model is partially identified in that he is only interested in identifying responses to the monetary policy shocks. Thus, rather than having to identify B_0^{-1}, it suffices to identify the vector b, which represents the structural responses of the VAR model variables to a monetary policy shock. Given $(A, P = \text{chol}(\Sigma_u))$, Uhlig proposes to use a simplex algorithm to solve the problem

$$b = \underset{b=Pq}{\arg\min}\, \Psi(b),$$

where the penalty function is defined as

$$\Psi(b) = \sum_k \sum_{h=0}^{H} f\left(\iota_k \frac{\theta_{k,h}(b)}{\sigma_k}\right),$$

with the first sum being over $k \in \{\text{GDP deflator, commodity price index, nonborrowed reserves, federal funds rate}\}$. The parameter σ_k is the standard deviation of the growth rate of variable k, $\theta_{k,h}$ is the structural response of variable k at horizon h to a monetary policy shock, $\iota_k = -1$ for the federal funds rate and $\iota_k = 1$ otherwise, and

$$f(x) = \begin{cases} x & \text{if } x < 0 \\ 100 \times x & \text{if } x \geq 0 \end{cases}.$$

Note that the sign of the penalty is flipped for the federal funds rate, reflecting the nature of the sign restrictions in Uhlig's model. The rescaling by σ_k serves

to make the deviations across structural impulse responses more comparable to one another.

It is worth pointing out that Uhlig's penalty function never directly imposes the sign restrictions, although it is very unlikely that these restrictions would not hold in the model that minimizes the penalty function. Another difference from the conventional approach is that this algorithm, by construction, will produce solutions even when the conventional approach yields the empty set, because the identifying sign restrictions do not have to be satisfied at the minimum of the penalty function. In the latter case, the penalty function given (A, P) will find the structural model that comes closest to satisfying the sign restrictions.

The particular form of the penalty function originally proposed by Uhlig (2005) as an alternative to the pure sign restriction approach has evolved to include many other loss functions. Examples include Mountford and Uhlig (2009) and Beaudry, Nam, and Wang (2012). The latter two studies postulate additional short-run exclusion restrictions, however, and hence do not belong into the current section. Penalty function approaches also play an important role in studies of forward-looking behavior (see, e.g., Barsky and Sims 2011).

Critiques of Penalty Function Estimators. It is important to differentiate this approach from earlier work by Faust (1998) that also involved penalty functions. Faust applies a penalty function to select among many candidate monetary policy VAR models identified by sign restrictions the model that explains most of the variation in real output growth at a horizon of nine years. Both Faust and Uhlig are interested in measuring the effects of monetary policy shocks on the variability of real output growth, given a set of admissible structural models. Moreover, both use a penalty function. It may seem that these approaches are closely related, therefore, but there are important differences.

Faust uses a penalty function to assess the best-case scenario relative to the economic hypothesis of interest, conditional on the sign restrictions having been imposed. The result is best thought of as providing an answer to the question of whether some outcome is possible rather than whether it is true or whether it is the most likely outcome. In contrast, Uhlig uses the penalty function to select the best model from the point of view of the economic hypothesis of interest and treats it like a point estimate. Moreover, his penalty function in practice involves an additional layer of identifying restrictions that all admissible structural models must satisfy. In Uhlig's words, "one is, in effect, imposing somewhat more than just sign restrictions" (see Uhlig 2005, p. 414). Finally, whereas Faust looks at the set of admissible structural models satisfying only the sign restrictions, Uhlig also considers models that come close to satisfying the restrictions implied by the penalty function. As a result, one would expect the individual structural impulse response estimates selected by this procedure to be different from the original estimates in the admissible

set. Clearly, this approach can only be recommended if we are comfortable with all the identifying restrictions implied by the use of a penalty function.

The problem is that the additional restrictions usually are not very transparent. Uhlig (2005) is aware of this concern and cautions that the pure sign restriction approach is cleaner and more appealing because it only imposes weak prior beliefs about the signs of impulse responses. Many applied users of penalty functions, however, seem unaware of the fact that they are imposing additional implicit identifying assumptions or, at any rate, have no idea what additional restrictions may be implied by the penalty function. Indeed, making these assumptions explicit can be prohibitively difficult.

In some cases the penalty function approach has been shown to imply additional unintended or at least unanticipated sign restrictions. Intuitively, the problem is that by choosing a particular orthogonal matrix that minimizes the response of one of the model variables to a given shock, we end up biasing the responses of other variables to the same shock. As a result, the penalty function involves additional restrictions on variables that are seemingly unrestricted (see Arias, Rubio-Ramírez, and Waggoner 2013). For example, Caldara and Kamps (2017) demonstrate that the penalty function used in Mountford and Uhlig (2009) amounts to imposing an additional sign restriction on the output response to a tax increase. A more detailed analysis of this example and the related work of Beaudry, Nam, and Wang (2012) can be found in Arias, Rubio-Ramírez, and Waggoner (2013), who explicitly trace the additional sign restrictions to the fact that the penalty functions in question reward structural models with large responses of some variables to some structural shocks.

These examples cast doubt on the use of penalty function estimators more generally. On the one hand, the use of a penalty function clearly cannot be recommended unless we understand its implications for identification. On the other hand, if we do, and if the implied additional restrictions involve sign restrictions as in the examples above, these sign restrictions should have been imposed directly, with the penalty function only serving to select a unique structural model.

Inference on Penalty Function Estimators. There are also questions about how to conduct valid inference for penalty function estimators. Uhlig (2005) proposes to apply the penalty function to the subset of sign-identified structural models obtained by exploring many solutions for Q, conditional on a given draw of the posterior for (A, P). This means that for each draw from the reduced-form posterior, the procedure will return just one estimate of the structural model. By repeating this procedure for many draws from the reduced-form posterior, one can approximate the posterior distribution of the penalty function estimator. Uhlig proposes to summarize this posterior by constructing posterior median response functions and pointwise posterior quantile bands.

There are three concerns with this approach. One concern is that, as noted earlier, the use of median response functions and pointwise error bands can be misleading. A second concern is that the penalty function approach tends to understate the posterior uncertainty (relative to the correct measure of uncertainty based on the explicit sign restrictions only) to the extent that it implicitly imposes additional identifying structure. The third concern is that measures of the posterior uncertainty of estimates obtained by the penalty function approach understate the uncertainty because they condition on one possible choice of the rotation matrix ignoring the existence of other solutions that satisfy the identifying restrictions (see Arias, Rubio-Ramírez, and Waggoner 2013).

Alternative Uses of Penalty Functions. None of these concerns arises if a penalty function is applied to choose among the models contained in the admissible set, as in Faust (1998). In that case we first generate many draws from the reduced-form posterior for (A, P) and, conditional on each of these draws, generate many draws of rotation matrices Q. Only the structural model solutions that satisfy the sign restrictions are retained. The resulting admissible set simultaneously accounts for estimation uncertainty and identification uncertainty. Applying a penalty function to choose among the models in this set will generate a unique structural model.

For example, Faust (1998) chooses the structural VAR model that maximizes the share of the variability of real output growth explained by the monetary policy shock. Similarly, Francis, Owyang, Roush, and DiCecio (2014) identify a technology shock as the shock that satisfies suitable sign restrictions and maximizes the forecast error variance share in labor productivity at some finite horizon. Unlike Uhlig (2005), they first impose the sign restrictions and then search for the structural model in the identified set that maximizes the forecast error variance.

Another example is Kilian and Lee (2014) who use an external estimate of the price elasticity of oil demand to select the structural oil market VAR model in the identified set that comes closest to implying a price elasticity of oil demand with that value. Rather than just establishing a best-case or worst-case scenario, as in Faust (1998), their analysis uses an external estimate of a model parameter to determine the best-fitting structural model within the admissible set.

13.6.5 Using Historical Information to Narrow the Set of Admissible Models

A very different approach to narrowing down the set of admissible models is to discriminate among these models based on their historical decompositions.

A case in point is Kilian and Murphy (2014) who utilize extraneous information about the role of oil supply shocks and speculative oil demand shocks during specific historical episodes to judge the economic plausibility of alternative admissible models. For example, there is extraneous information from oil industry sources about a surge in speculative demand in the second half of 1979, allowing us to discard models that do not replicate this feature. Likewise we know that the spike in the price of oil in 1990, following the invasion of Kuwait, must be related at least in part to an oil supply disruption and was not caused by a booming economy. We also know that there was an important shift in oil price expectations that must have been reflected in increased speculative demand and higher oil prices in 1990.

This approach amounts to imposing additional identifying restrictions that can only be imposed on the historical decompositions rather than on the structural impulse responses. Unlike the penalty function approach, this approach typically does not produce a single admissible model, but it may be used to narrow down the admissible set considerably. Of course, its application is limited to settings in which the researcher has accurate information about the quantitative importance of a given structural shock during a specific historical episode.

This approach has recently been formalized by Antolín-Diaz and Rubio-Ramírez (2016). Building on the framework of Rubio-Ramírez, Waggoner, and Zha (2010), they show how both the sign of shocks occurring on specific dates and the sign of the cumulative contribution of structural shocks during specific episodes may be restricted when estimating the sign-identified VAR models. For example, one may impose the restriction that the contribution of one shock is larger in absolute magnitude than that of some other shock (or possibly of all other structural shocks combined) during a given episode. These additional identifying restrictions based on extraneous evidence may considerably reduce the range of model solutions consistent with the data and tend to tighten the credible sets. Antolín-Diaz and Rubio-Ramírez (2016) apply this approach to the oil market VAR model of Kilian and Murphy (2012) and the semistructural VAR model of monetary policy of Uhlig (2005).

13.7 The Role of the Prior for the Rotation Matrix

The reason why classical estimation methods are inherently uninformative about which of the admissible structural models is most likely is that the likelihood is flat with respect to the choice of the rotation matrix. An obvious question is how Bayesian methods are seemingly able to overcome this problem. From a mechanical point of view, the answer is that they rely on a prior distribution for the rotation matrix Q. This prior also is known as a Haar prior in the literature. Applying the Givens and Householder transformations to generate draws for Q is equivalent to generating draws from this prior. Although

the marginal prior distribution of Q is uniform in the Haar space, as noted earlier, the implied priors for the impact multiplier matrix used in constructing the structural impulse responses are clearly informative. This result follows from the fact that the impact responses are a weighted average of the elements of Q which themselves are not uniform. The resulting informative prior for the structural impulse responses is not based on economic information, however, and there is no way for the data to overrule this prior even asymptotically because the likelihood does not depend on Q. This fact raises obvious concerns about the conventional Bayesian approach to estimating structural VAR models subject to sign restrictions.

An important practical question is to what extent the posterior distribution of the estimated structural VAR models, Θ, depends on the prior for Q, as opposed to the data. On the one hand, one certainly can construct examples in which the posterior essentially mirrors the prior. On the other hand, given that Inoue and Kilian (2013) evaluate the posterior of Θ, which includes structural impulse responses beyond the impact period, it is possible that much of the posterior information, which depends in part on A and P, may come from the data rather than the prior on Q. This is an empirical question. If the location and the concentration about the mode of the posterior distribution differed substantially from the location and concentration of the prior distribution of Θ, this outcome would increase our confidence in the posterior of the sign-identified model. If not, the outcome would cast doubt on the results from sign-identified VAR models. Of course, this diagnostic only reveals the extent to which the prior on Q affects inference about the structural impulse responses without addressing the root cause of this problem.

The ultimate cause of this problem is that the traditional approach does not specify the prior directly on the object of interest, which are the structural impulse responses, but on the model parameters on which the structural impulse responses depend. One response to this problem has been the development of the frequentist approaches to inference in sign-identified models reviewed in Section 13.6.1. These approaches have their own limitations, however. Another response has been to modify the way Bayesian estimation and inference is conducted.

13.7.1 An Approach Based on Explicit Bayesian Priors for B_0

Baumeister and Hamilton (2015a) formally demonstrate that the marginal prior distribution for Q is informative about the structural model parameters even asymptotically. They show that the nature of this implicit prior distribution varies with the dimension K of the VAR model. For example, in a standard bivariate model of price and quantity driven by demand and supply shocks, in which sign restrictions on the impact responses are the only identifying assumptions, the price elasticities of demand and supply, which can be shown

to be linear functions of selected elements of B_0, have a truncated Cauchy distribution under conventional prior specifications for sign-identified structural VAR models. If the reduced-form residuals are positively correlated, the model a priori allows any value of the price elasticity of demand but restricts the price elasticity of supply to fall within a certain interval. With negatively correlated errors, in contrast, the elasticity of supply could be any positive number, whereas the elasticity of demand is restricted to fall within a particular interval. Thus the choice of the Haar prior in this example is anything but innocuous. Although such prior restrictions on the range of the price elasticities need not arise in larger-dimensional models, as shown in Kilian and Murphy (2014), there is no doubt that the use of the Haar prior may affect the posterior of the structural impulse responses and the implied price elasticities in unknown and possibly undesirable ways.

Baumeister and Hamilton's proposed solution is to specify the prior directly on the elements of B_0, generalizing the approach of Sims and Zha (1998). They impose the restriction that the marginal priors for each element of B_0 are independent of one another, facilitating the task of specifying the joint prior. The marginal priors on the elements of B_0 may be uninformative (flat or diffuse) or deliberately informative. One obvious drawback of this proposal is that any prior on the elements of B_0 implies an informative prior on B_0^{-1}. The nature of that prior is not made explicit and need not coincide with the prior views of the researcher about the structural impulse responses. Thus, unless the case can be made that the elements of B_0 are the object of ultimate interest rather than the structural impulse responses, this approach suffers from exactly the same problem as the traditional approach it seeks to replace. Put differently, Baumeister and Hamilton (2015a) do not propose a solution to the problem of specifying a prior directly on B_0^{-1} (or more generally on all structural impulse responses), but they are proposing a solution to the problem of modeling the parameters of the structural VAR model.

There are examples where their approach is appealing, as illustrated by the bivariate empirical example in Baumeister and Hamilton (2015a). For many applications of sign-identified VAR models in the literature, however, this approach is problematic. This point is illustrated by the additional empirical examples provided in Baumeister and Hamilton (2015c).

Priors on Elasticities. Baumeister and Hamilton (2015c) focus on oil market VAR models including the model of Kilian and Murphy (2012). The central premise of Baumeister and Hamilton is that the price elasticities of demand and supply can be written as linear functions of the elements of B_0, allowing them to impose priors on these elasticities. This approach makes sense provided that at each point in time the production and consumption of the commodity in question coincide, as assumed in the standard bivariate model of prices and quantities. That assumption does not hold for storable commodities such as

crude oil, however. In the latter case, production equals consumption plus the change in inventories. As a result, the price elasticity of oil demand no longer is a linear function of the elements of B_0 (see Kilian and Murphy 2014).

This fact not only makes it impossible to impose economically motivated priors on this elasticity by means of priors on selected elements of B_0, but it also invalidates Baumeister and Hamilton's maintained assumption that the marginal priors on B_0 are independent. The price elasticity of oil demand, properly defined to account for the role of oil inventories, depends on both the response of production and the response of inventories to a change in the real price of oil triggered by an exogenous shift in the oil supply curve. Hence, a prior on this price elasticity by construction has implications for the prior of more than one element of B_0, violating the assumption of independence. In short, Baumeister and Hamilton's approach is not suitable for this class of problems.

In addition, this fact matters for the interpretation of the results. Baumeister and Hamilton (2015c) suggest that oil market VAR models such as Kilian and Murphy (2012) or Kilian (2009) that impose tight upper bounds on the price elasticity of oil supply are fundamentally flawed because they imply unrealistically large price elasticities of oil demand. That conclusion, however, is based on Baumeister and Hamilton misinterpreting the coefficients of B_0 in the model of Kilian and Murphy (2012) as implying an estimate of the price elasticity of oil demand. In fact, that model (like the earlier model of Kilian 2009) is not designed to infer the price elasticity of oil demand. If the structural oil market VAR model is suitably modified to allow proper estimation of this elasticity, as shown in Kilian and Murphy (2014), price elasticities of oil supply close to zero are perfectly consistent with reasonably low price elasticities of oil demand.

Informative versus Uninformative Priors. Baumeister and Hamilton suggest that we embrace the fact that priors in sign-identified VAR models are informative and make this information content explicit. An obvious concern with intentionally specifying informative priors for the elements of B_0 in sign-identified models is that it is not clear why a prior that merely reflects the personal views of the user should be of wider interest to other economists. In fact, an unsympathetic observer may view the use of subjective priors as an attempt to circumvent the lack of identification by simply imposing the answer. Thus, at best this approach may provide an opportunity for illustrating how alternative subjective priors affect the posterior.

Baumeister and Hamilton are aware of this concern and make the case that imposing informative priors makes sense when there is extraneous identifying information on the structural model parameters that has not been utilized in existing research. For example, there may be extraneous information about the value of elasticity parameters from microeconomic studies that can be imposed

in estimation. Given the uncertainty about such estimates, it makes sense to think of this information as being stochastic with the prior parameterizing our degree of confidence in these estimates. Unlike conventional subjective priors, priors based on extraneous evidence can be viewed as objective in that there should be agreement among economists on this prior information.

This approach clearly is appealing at first sight. As the oil market VAR examples in Baumeister and Hamilton (2015c) illustrate, however, it can be difficult to find such identifying information beyond restrictions such as elasticity bounds that have already been imposed in the existing literature (see Kilian and Murphy 2014; Kilian and Lee 2014). Moreover, it is not clear how exactly to map extraneous evidence about elasticities into priors without introducing subjective elements.

What Do We Learn from Extraneous Elasticity Estimates? It is important to be precise about what we learn from extraneous evidence about the price elasticities of oil demand and oil supply. These parameters play a central role in oil market VAR models. Kilian and Murphy (2012) demonstrate that the price elasticity of oil supply is crucial in determining the relative importance of oil demand and oil supply shocks in explaining oil price fluctuations. If that elasticity is close to zero, one obtains the by now standard result that oil demand shocks have been the primary driver of oil price fluctuations all the way back to the 1970s, with rare exceptions such as 1990 and to some extent 2014/15.

The rationale for a price elasticity of oil supply close to zero in Kilian and Murphy (2012) and related studies is threefold: (a) Narrative evidence on decisions by major oil producers such as Saudi Arabia shows a reluctance to respond to price fluctuations driven by oil demand (see Kilian 2009); (b) microeconometric evidence based on production decisions by U.S. oil producers suggests elasticity estimates of 0.0009 (e.g., Anderson, Kellogg, and Salant 2016); (c) economic theory shows that in equilibrium, oil producers do not respond to oil price fluctuations caused by oil demand shocks in the short run; all the adjustment works through investment (see Anderson, Kellogg, and Salant 2016).

Baumeister and Hamilton depart from this consensus and make the case for priors that put more weight on lower price elasticities of oil demand and higher price elasticities of oil supply than found in earlier oil market VAR studies. Such a change in the prior has far-reaching implications because it is expected to improve the ability of oil supply shocks to explain fluctuations in the real price of oil at the expense of oil demand shocks. In support of their prior, Baumeister and Hamilton appeal to a range of estimates of price elasticities of oil demand and oil supply in the existing literature, some of which date back to the 1970s and 1980s.

It may seem tempting simply to aggregate all existing extraneous elasticity estimates, but that approach would be misleading because not all estimates in the literature are equally credible. Although there are numerous extraneous microeconometric estimates of the short-run price elasticity of oil demand in the literature, many of these estimates are questionable. Not only do traditional estimates of this elasticity ignore the storability of crude oil, but they are based on OLS reduced-form regressions of quantity on price and hence suffer from simultaneous equation bias. Econometric theory implies that these elasticity estimates are biased toward zero and are not economically meaningful. More recent estimates of the price elasticity of oil demand from structural econometric models are invariably larger, as noted by Kilian and Murphy (2014). This conclusion is consistent with recent IV estimates of the short-run price elasticity of gasoline demand, which in turn have been externally validated (see Coglianese, Davis, Kilian, and Stock 2017). For example, the estimate of the one-month price elasticity of oil demand of -0.25 in Kilian and Murphy (2014) is much higher than the modal value of -0.1 in Baumeister and Hamilton's baseline prior specification. Likewise, for the short-run price elasticity of oil supply, there are few credible microeconometric estimates, but the available empirical evidence, anecdotal evidence, case studies, and economic theory all point to an elasticity much closer to zero than Baumeister and Hamilton's modal value of 0.1. Thus, there is no evidence that the existing literature on oil market VAR models has ignored important identifying information, but there are questions about the empirical support for the elasticity priors proposed by Baumeister and Hamilton.

How Do We Translate Extraneous Elasticity Estimates into a Prior?
Even if one can agree on a reasonable prior mean or mode for these elasticities, the problem remains of how to translate the extraneous information into prior densities for the elasticities that reflect more than introspection. Say, for example, that some earlier study reports that the short-run price elasticity of oil supply is 0.02. This does not tell us what the functional form of the prior density should be. Nor does it tell us what the dispersion of the prior density should be. In fact, even if one had multiple credible estimates of the price elasticity of supply, say 0.05 and 0.02, this would not suffice to specify the prior density any more than one estimate. In other words, the elasticity priors specified by Baumeister and Hamilton in the end come from introspection rather than extraneous information. Perhaps for this reason, Baumeister and Hamilton (2015c) appear to favor priors intended to be fairly diffuse. In the end, they seek to establish that their main results are robust to alternative prior specifications for B_0 rather than relying on additional extraneous evidence.

Incompatibility with the Identifying Restrictions Used in Earlier Studies. Baumeister and Hamilton specify independent marginal priors on each element

of B_0. These priors may take the form of a truncated uniform or a t distribution, for example. As discussed earlier, in applied work, it is determined by the economic model of interest whether a user imposes identifying restrictions on the elements of B_0 or B_0^{-1}. These representations in general are not exchangeable.

Thus, when Baumeister and Hamilton (2015c) attempt to illustrate the benefits of their approach to estimation and inference in the context of the oil market VAR model of Kilian and Murphy (2012), for example, they do not rely on the same economic structure and on the same identifying assumptions as the original study. Their analysis both ignores some of the identifying sign restrictions imposed by Kilian and Murphy (2012) and imposes additional exclusion restrictions on B_0 not found in the original specification. Specifically, Baumeister and Hamilton postulate that global real economic activity only affects oil production contemporaneously through its effect on the real price of oil and that global oil production does not enter directly in the global real economic activity equation. Hence, the model considered by Baumeister and Hamilton is at best similar to the original model, making it difficult to attribute differences in the estimates to the choice of the prior. Put differently, the point is not that Baumeister and Hamilton chose not to replicate the original identifying assumptions in Kilian and Murphy (2012) because they disgree with their economic content, but that their methodology does not allow them to incorporate these assumptions.

More generally, the alternative approach proposed by Baumeister and Hamilton does not allow for any dynamic sign restrictions or cross-equation restrictions of the type utilized by many earlier oil market VAR models. Baumeister and Hamilton (2015a) argue that such restrictions are irrelevant for identification, but that conclusion only holds under the particular loss function they adopt.

In short, the objective of the analysis in Baumeister and Hamilton is not to provide an improved way of estimating the sign-identfied models proposed by other researchers; rather it is to replace the structural models and identifying restrictions used by those researchers by alternative models and assumptions. Whether these alternative models are more or less economically appealing than the original models has to be decided on a case-by-case basis.

Choice of the Loss Function. Any comparisons with the results in the existing literature are further complicated by the fact that Baumeister and Hamilton choose to rely on a specific loss function that allows them to effectively use the median response function as the point estimate of the structural impulse response function. This choice of the loss function is controversial. Median response functions suffer from the shortcomings we discussed earlier and are not informative about the questions most applied users care about. Whether the

prior for the elements of B_0^{-1} is implicit or explicit does not affect this conclusion. As illustrated in Inoue and Kilian (2013), for a given prior, evaluating the set of admissible models based on the median response function rather than the responses of the most likely structural model may affect not only the magnitude but even the sign of the structural responses.

Thus, it is difficult to know whether any differences between the estimates reported in Baumeister and Hamilton (2015c) and in the original studies are driven by changes in the model specification, in the prior, in the identifying assumptions, or in the loss function. Notwithstanding these caveats, Baumeister and Hamilton (2015c) end up confirming many of the substantive conclusions of earlier oil market studies, suggesting that the prior for Q embodied in conventional Bayesian estimates of sign-identified structural VAR models is not overly informative in these applications.

13.7.2 An Approach Based on Explicit Bayesian Priors for the Structural Impulse Responses

A closely related alternative approach that allows the user to impose priors directly on elements of B_0^{-1} and on the shapes and smoothness of the structural impulse response functions has been proposed by Plagborg-Møller (2016). Let w_t be a K-dimensional vector of structural shocks. In practice, w_t is typically modeled as iid Gaussian. The central idea is to estimate the structural moving average representation,

$$y_t = \Theta(L)w_t,$$

up to the maximum horizon of interest directly rather than estimating the structural VAR representation,

$$B(L)y_t = w_t,$$

first and then inverting that representation, as proposed by Baumeister and Hamilton (2015a). Plagborg-Møller proposes to impose a multivariate prior directly on the first $H + 1$ coefficient matrices of $\Theta(L)$, including the structural impact multiplier matrix B_0^{-1} and the structural impulse responses at horizons $h = 1, \ldots, H$. As in Baumeister and Hamilton (2015a), this approach can be used to explicitly allow for identification uncertainty or it may be used to impose a degenerate prior on certain elements of B_0^{-1}. Thus, traditional short-run and long-run exclusion restrictions as well as static and dynamic sign restrictions on structural impulse responses are covered as special cases. Short-run exclusion restrictions may be imposed by dropping the elements in question from the set of structural impulse responses. Long-run exclusion restrictions may be imposed as a constraint in evaluating the likelihood. Sign restrictions may be imposed by restricting the parameter space of the

structural impulse responses to the subspace where the inequality restrictions hold.

One potential advantage of this approach compared with Baumeister and Hamilton (2015a) is that it is in line with most applied studies in seeking to restrict the structural impulse responses rather than the parameters of the structural model. It also allows the user to impose shape restrictions on the structural impulse response functions. Such restrictions can be incorporated as constraints in the evaluation of the likelihood. Moreover, the smoothness of the structural impulse response functions can be controlled by the prior covariance across structural impulse responses.

The obvious cost is that the parameters of the Wold reduced-form MA representation are no longer pinned down by the data alone. To conduct posterior inference about the structural impulse responses, Plagborg-Møller (2016) develops a simulation algorithm that exploits the Whittle (1953) likelihood approximation. He shows how this posterior may be approximated by a Markov Chain Monte Carlo method. He then derives the frequentist limit of this posterior distribution under weaker assumptions than Moon and Schorfheide (2012). Under some simplifying assumptions including covariance stationarity of the data, it can be shown that the data will asymptotically pin down the coefficients of the reduced-form Wold representation, but the structural impulse responses in this framework are only set-identified, making their posterior distribution sensitive to the choice of the prior for B_0^{-1}. Plagborg-Møller follows Baumeister and Hamilton in embracing this fact, arguing for the imposition of explicitly informative priors in identifying the structural impulse responses.

The main challenge in implementing Plagborg-Møller's approach thus is the derivation of the priors. There are two distinct concerns. One is that much of what we learn from the posterior reflects the prior imposed in estimation. This is a bigger concern than in other approaches to sign-identified models, because imposing a prior on the first $H+1$ coefficient matrices of $\Theta(L)$ implicitly imposes an informative prior on the reduced-form representation of the model. Even if the effect of that prior on the posterior is negligible asymptotically, it will matter in finite samples.

The other concern is that coming up with universally accepted priors may be difficult. For example, it is not clear how to derive a nondogmatic prior density for the elements of B_0^{-1} in general. Nor is there empirical support for the claim that economists in the past have routinely failed to impose readily available identifying information. Plagborg-Møller shows by example that in some situations priors for the parameters of $\Theta(L)$ may be elicited from DSGE models, but in many applications that approach may not be feasible or desirable. Moreover, even if a particular economic model suggests a certain pattern of identifying restrictions, this identifying information may not be robust to changes in the specification of this economic model. Imposing a diffuse prior

on the dynamics of the structural impulse responses, in contrast, avoids this problem, but is likely to imply very wide credible sets.[5]

Another unresolved question is how to report the posterior evidence. Plagborg-Møller relies on posterior mean vectors for the structural responses and pointwise HPD error bands. He also reports selected elements of the shotgun trajectory plot. A useful generalization would be to derive the most likely structural model and the joint credible set from the posterior distribution, building on the analysis in Inoue and Kilian (2013).

In short, the approach taken by Plagborg-Møller (2016) provides a useful complement to the analysis in Baumeister and Hamilton (2015a) in that it considers identifying restrictions not covered by their analysis. At this point, we do not have much experience in making this alternative approach operational, however. It remains to be seen whether this proposal may be developed into a practically useful alternative to conventional sign-identified VAR models.

13.7.3 A Robust Bayesian Approach

A very different response to the problem of informative priors on Q in sign-identified structural VAR models is due to Giacomini and Kitagawa (2015). Whereas Baumeister and Hamilton (2015a) propose deriving prior densities that reflect the user's beliefs about the parameters of B_0 and Plagborg-Møller (2016) proposes imposing priors directly on the structural impulse responses (including the elements of B_0^{-1}), Giacomini and Kitagawa's proposal is to construct posterior bounds on B_0^{-1} (and more generally on the structural impulse responses) without taking a stand on the nature of the prior for Q. In other words, Giacomini and Kitagawa are not concerned with the economic content of the prior for Q. The only economic information used in their approach involves the imposition of sign restrictions (or other identifying restrictions) on the structural impulse responses.

Giacomini and Kitagawa show that the problem of constructing pointwise credible sets from the posterior of the structural responses derived under all possible priors for Q is equivalent to constructing pointwise posterior bounds on each of the structural impulse responses. Their procedure allows the structural VAR model to be partially identified (or underidentified) in that only a subset of the structural shocks is identified. In practice, Giacomini and Kitagawa's procedure only relies on a prior for the reduced-form parameters of the VAR model and on a set of identifying restrictions. One practical drawback of their procedure is that the construction of the bounds involves a complicated

[5] A closely related Bayesian approach has been developed by Barnichon and Matthes (2016). Unlike Plagborg-Møller, they impose additional structure on the coefficients of the structural MA representation to reduce the dimensionality of the estimation problem. Specifically, they approximate the impulse response functions with so-called Gaussian basis functions.

nonlinear optimization problem for each posterior draw. Moreover, in practice, the resulting credible sets are likely to be so wide as to be uninformative. Finally, as in frequentist approaches to inference in sign-identified models, the estimates can be difficult to interpret from an economic point of view.

13.7.4 An Agnostic Bayesian Approach

An alternative Bayesian approach was recently proposed by Arias, Rubio-Ramírez, and Waggoner (2015). Their objective is to ensure that the prior we use in applied work is agnostic in that it does not imply additional restrictions beyond the sign restrictions that the user wishes to impose. The analysis is based on the observation that we can represent a structural VAR model in terms of the parameters of the structural VAR model representation, $B_0, \ldots, B_p$, in terms of a set of structural impulse responses, Θ, or in terms of the representation (A, Σ_u, Q). Given A and P such that $PP' = \Sigma_u$, each value of the orthogonal matrix Q determines a particular structural model (see also Inoue and Kilian 2013). Arias et al. define a prior over the structural model representation (or over the structural impulse response representation) to be agnostic with respect to the identification if the prior density is invariant to the choice of $Q \in \mathcal{O}(K)$ and show that a prior is agnostic if and only if it is equivalent to a prior over (A, Σ_u, Q) that is flat over $Q \in \mathcal{O}(K)$. This result immediately implies that the conventional Bayesian approach to estimating models subject to sign restrictions discussed earlier is agnostic in this sense. It is important to note, however, that a prior being agnostic in this technical sense does not mean that the prior is not informative for the structural impulse responses. In particular, the implied prior for the structural impulse responses is not flat.

As an alternative to the agnostic prior, Arias et al. also consider the construction of jointly flat priors over the structural model representation or over the structural impulse response representation, respectively. Their representation of this problem is more general than that in Baumeister and Hamilton (2015a) in that it allows for the imposition of dynamic sign restrictions in addition to static sign restrictions. Arias et al. prove that if a prior is flat over either the structural model representation or the structural impulse response representation, then the equivalent of that prior written in terms of the representation (A, Σ_u, Q) is flat over $\mathcal{O}(K)$. The fact that the conventional prior discussed in Rubio-Ramírez, Waggoner, and Zha (2010) is informative about the structural model parameters and about the structural impulse response parameters thus does not result from the assumption of a flat prior over the rotation matrices Q; rather, it stems from their choice of the prior for the reduced-form parameters. This fact suggests that we need to change the reduced-form prior to ensure that the prior is flat in the desired dimension as well as agnostic.

Arias et al. propose to choose a conjugate prior within the class of Gaussian-inverse Wishart priors such that the prior is flat over either the structural model

representation or the structural impulse response representation, depending on the user's preferences. One can be flat in one of these dimensions but not in both. They show that a prior over the structural model representation is flat if and only if the equivalent prior over the representation (A, Σ_u, Q) equals

$$2^{\frac{-K(K+1)}{2}} |\det(\Sigma_u)|^{-\left(\frac{Kp}{2}+K+1\right)},$$

which implies a flat prior for Q on $\mathcal{O}(K)$ and a Gaussian-inverse Wishart prior of a particular form. Similarly, a prior over the structural impulse response representation is flat if and only if the equivalent prior over the representation (A, Σ_u, Q) equals

$$2^{\frac{-K(K+1)}{2}} |\det(\Sigma_u)|^{-\left(\frac{Kp}{2}-1\right)},$$

which implies a flat prior for Q on $\mathcal{O}(K)$ and a Gaussian-inverse Wishart prior of a different form.

13.7.5 A Non-Bayesian Approach

Finally, a very different approach has been pursued in Baumeister and Kilian (2016b) in the context of a global oil market VAR model. Their objective is to quantify the extent to which the decline in the price of oil in the second half of 2014 is explained by oil demand and oil supply shocks taking place after June 2014. For this purpose, Baumeister and Kilian develop a novel identification strategy based on recursive estimates of the reduced-form VAR representation. As in Kilian and Murphy (2014), the vector of VAR model variables includes global oil production, global real economic activity, global crude oil inventories, and the real price of crude oil. At each point in time, the VAR model generates a set of reduced-form prediction errors for each of the model variables obtained by comparing the one-step ahead model predictions with the subsequent realizations of the model variables. These prediction errors by construction reflect the oil demand and oil supply shocks in the underlying structural VAR model.

If a given set of prediction errors is driven primarily by one of the structural oil demand and oil supply shocks, we can use identifying information about the signs and relative magnitudes of the prediction errors to infer which of the structural shocks can explain the observed pattern of prediction errors. For example, a negative oil supply shock would be associated with a large positive prediction error in the real price of oil and a large negative prediction error in oil production, but a small negative prediction error for real activity and a small negative prediction error for oil inventories. Whether a prediction error for oil production is large can be judged based on the existing literature on large oil supply shocks (see, e.g., Hamilton 2003). Likewise, prediction errors for oil inventories can be expressed relative to the stock of oil inventories

to obtain a measure of their magnitude. The corresponding patterns implied by flow demand or speculative demand shocks may be derived in a similar manner.

This identifying information allows one to assess whether the pattern of prediction errors for a given month is consistent with any one of the structural shocks. It is possible, of course, for the prediction errors to reflect a multitude of structural shocks of large magnitude. In the latter case, the pattern of prediction errors will not fit the patterns implied by any one structural shock, and the shocks cannot be identified. It is also possible that the shocks in a given month are so small that the prediction error for the real price of oil is negligible, in which case the question of identification is moot.

In their particular application, Baumeister and Kilian (2016b) show that between July 2014 and December 2014 there were only two large oil price prediction errors, one in July and one in December. The first prediction error was unambiguously associated with a reduction in speculative demand driven by a shift in expectations about the future price of oil, and the second prediction error was unambiguously driven by a reduction in flow demand, reflecting an unexpected weakening of the global economy. In contrast, the alternative hypothesis of a shift in oil price expectations in December 2014, triggered by the OPEC announcement of late November, can be ruled out on the basis of the pattern of prediction errors.

The advantage of this approach to identification is that it does not require any prior for Q, and, in fact, does not require the user to employ a Bayesian approach to estimating the VAR model, yet it may allow one to quantify numerically the extent to which structural shocks shifted the real price of oil. Its most important disadvantage is that it does not generate numerical estimates of the structural shocks. The effect of a structural shock can only be judged informally by comparing the path of model predictions after the shock has occurred to the path predicted before this shock occurred.

13.8 Examples of Models Identified by Sign Restrictions

13.8.1 A Small-Scale Macroeconomic Model

In Chapter 8, we examined a stylized quarterly semistructural model of monetary policy, in which the monetary policy shock was identified by exclusion restrictions that ruled out contemporaneous feedback from the policy shock to the variables ordered above the interest rate. These identifying assumptions are neither credible in general nor consistent with the implications of most general equilibrium models. Sign restrictions provide a natural alternative.

As discussed in Fry and Pagan (2011), we may postulate instead a positive demand shock that on impact raises economic growth (Δgdp), inflation (π), and the interest rate (i); a positive cost-push (or negative supply)

shock that on impact lowers growth but raises inflation and the interest rate; and a contractionary monetary policy shock that on impact raises the interest rate but lowers inflation and growth. Formally these sign restrictions can be represented as

$$\begin{pmatrix} u_t^{\Delta gdp} \\ u_t^{\pi} \\ u_t^{i} \end{pmatrix} = \begin{bmatrix} + & - & - \\ + & + & - \\ + & + & + \end{bmatrix} \begin{pmatrix} w_t^{\text{AD}} \\ w_t^{\text{AS}} \\ w_t^{\text{monetary policy}} \end{pmatrix}.$$

13.8.2 A Slightly Larger Macroeconomic Model

Peersman (2005) postulates a model of the U.S. economy based on a covariance stationary structural VAR model for the percent change in the nominal price of oil (Δo), real output growth (Δq), consumer price inflation (Δp), and the short-term nominal interest rate (i). Cointegration among the variables in levels has been ruled out. The real price of oil is implicitly assumed to be $I(1)$ so that Δo is $I(0)$. The vector of structural shocks includes a nominal oil price shock, a domestic aggregate supply shock, a domestic aggregate demand shock, and a domestic monetary policy shock. Identification involves sign restrictions of the form

$$\begin{pmatrix} u_t^{\Delta o} \\ u_t^{\Delta q} \\ u_t^{\Delta p} \\ u_t^{i} \end{pmatrix} = \begin{bmatrix} + & * & + & - \\ - & + & + & - \\ + & - & + & - \\ + & - & + & + \end{bmatrix} \begin{pmatrix} w_t^{\text{oil price}} \\ w_t^{\text{AS}} \\ w_t^{\text{AD}} \\ w_t^{\text{monetary policy}} \end{pmatrix},$$

where * stands for an unrestricted element.

An unexpected monetary policy tightening is associated with higher interest rates but lower real output and lower inflation on impact. The impact effect on the nominal price of oil is negative, consistent with the response of the price level. A positive domestic aggregate demand shock is associated with positive impact responses of all model variables. A positive oil price shock is interpreted as a negative domestic aggregate supply shock. It raises the price of oil, raises inflation, lowers real output, and raises the interest rate. A positive domestic aggregate supply shock, in contrast, shows the opposite signs, except that its effect on the nominal price of oil is treated as unknown because of the conflicting signs of the real and nominal effects of such a shock. On the one hand, one would expect the nominal price to increase because of higher real demand for oil; on the other hand, one would expect it to fall because of the deflationary effects of positive domestic aggregate supply shocks on dollar-denominated prices.

On the basis of the sign restrictions alone, it cannot be ruled out that a positive oil price shock may be observationally equivalent to a negative domestic

aggregate supply shock because both shocks would have in common the same sign pattern if the negative aggregate supply shock were to lower the nominal price of oil, which is not precluded by the assumptions so far. Because the oil price shock and the aggregate supply shock are not individually identified, Peersman assumes that of these shocks the oil price shock is the shock with the larger impact effect on the nominal price of oil. This restriction may be imposed as an additional inequality restriction enforcing that one response is larger than the other. The sign restrictions on the price level and real output are imposed not only on impact but for the first four quarters.

13.8.3 A Model of Unemployment and Vacancies

Fujita (2011) proposes a quarterly structural VAR model including the job separation rate, the job finding rate, and the number of vacancies in the United States. The number of vacancies is measured by the number of help-wanted advertisements in newspapers. All data are detrended to remove low-frequency variation not explained by economic theory. The model is partially identified in that only the responses to a shock in the profitability of the employment relationship is identified. A positive shock to profitability is assumed to increase vacancies on impact, resulting in a sign restriction on B_0^{-1}. A positive shock to profitability also is assumed to cause declines in unemployment for the first two quarters. Although changes in unemployment are not included in the reduced-form model, the unemployment response may be inferred from the responses of the transition rates via the implied gross job flows (for details see Fujita 2011). Thus, the second identifying assumption imposes nonlinear restrictions on the responses of the two transition rates. This very simple model is consistent with a wide range of Mortensen-Pissarides style search and matching models with and without endogenous job separation decisions.

13.8.4 An Extended Model of Unemployment and Vacancies

Fujita (2011) also considers an extended model designed to differentiate between two different sources of shocks to profitability. The model includes inflation and productivity growth in addition to job finding and job separation rates and vacancies. The model is again partially identified, but rather than identifying one shock to profitability, Fujita identifies separately a demand shock and a productivity shock. It is assumed that a positive demand shock raises the price level for the first four quarters, causes vacancies to rise on impact and the change in unemployment to be negative for the first two quarters. A positive technology shock is assumed to increase labor productivity for the first 20 quarters, to lower the price level for the first four quarters, and to raise unemployment. The change in unemployment again is defined on the basis of the responses of the rates of transition.

13.8.5 A Model of Technology Shocks

Dedola and Neri (2007) propose a quarterly structural VAR of the U.S. macroeconomy designed to study the responses of the economy to a positive technology shock. The variables in the model are the log of labor productivity, real wages, per capita hours worked, per capita real investment, per capita real consumption, the GDP deflator rate of inflation, and the short-term interest rate.

The sign restrictions are explicitly derived from a representative DSGE model. A positive technology shock increases labor productivity for the first 20 quarters, investment and output for the first 10 quarters, real wages for all quarters between the third and the twentieth quarter, and consumption for the first 5 quarters. The responses of hours, inflation, and the short-term interest rate are left unrestricted.

13.8.6 A Model of Exchange Rate Responses to Monetary Policy Shocks

Scholl and Uhlig (2008) propose a model of the response of the monthly nominal exchange rate to monetary policy shocks that is identified by static and dynamic sign restrictions as well as shape restrictions on selected impulse responses. They study the relationship between the United States and other economies, one economy at a time. The reduced-form is identical to the recursively identified model of Eichenbaum and Evans (1995). The model includes seven variables: U.S. and foreign industrial production, U.S. and foreign short-term interest rates, the U.S. price level, the ratio of nonborrowed reserves to total reserves in the United States, and the bilateral nominal dollar exchange rate in dollars per foreign currency. All variables but the interest rates are in logs.

Scholl and Uhlig (2008) consider two alternative sets of identifying assumptions. The first identification scheme postulates that an unanticipated monetary tightening in the U.S. lowers the price level and the ratio of nonborrowed to total reserves for the first year, while raising the interest rate for the first year.

The second identification scheme postulates that an unanticipated monetary tightening in the U.S. lowers the price level and the ratio of nonborrowed to total reserves for the first half year, while raising the interest rate for the first half year. It also adds the restrictions that the response of the U.S. interest rate exceeds the response of the foreign interest rate for the first half year. Finally, it imposes that the impact response of the nominal exchange rate is negative, and it imposes the absence of delayed overshooting in the exchange rate. The latter identifying assumption implies a shape restriction on the response function of the exchange rate that can be written as

$$s_0 < 0, \ \left|s_j\right| < \left|s_{j-1}\right|, \ |s_l| < |s_0| \ \text{ for } j = 0, 1, 2 \text{ and } l = j+1, \ldots, 23,$$

where s_j denotes the response of the exchange rate after j periods to an unexpected monetary policy tightening at date 0. Delayed overshooting refers to a situation in which a contractionary monetary policy shock causes a gradual appreciation of the exchange rate, followed by a gradual depreciation. In other words, the response reaches its maximum only with a delay. This pattern is inconsistent with traditional rational expectations open-economy sticky price models such as Dornbusch (1976) and is ruled out by the shape restriction above.

13.8.7 A Medium-Scale Macroeconomic Model

Canova and Paustian (2011) is an example of a medium-scale structural VAR model of the U.S. economy that simultaneously identifies many shocks. In Chapter 6 we already discussed in some detail the relationship between VAR models and DSGE models. Canova and Paustian first examine impulse responses for a wide range of medium-scale DSGE models. They document robust sign patterns for a range of structural shocks in these theoretical models and then proceed to impose these sign patterns as identifying assumptions on structural VAR models.

One implication of their analysis is that dynamic sign restrictions tend not to be robust across alternative theoretical models and hence should be avoided when fitting macroeconomic models. Another implication is that the choice of model variables matters. In VAR models with few variables it may not be possible to discriminate between the responses to alternative structural shocks because different structural shocks imply the same sign pattern in the responses of variables that are included in the VAR model. Usually, this problem may be overcome by judiciously adding variables, the responses of which help discriminate between alternative shocks.

The baseline model in Canova and Paustian (2011) includes five variables (nominal interest rate, real wage, inflation, real output, and hours worked) and identifies four structural shocks (markup shock, monetary shock, taste shock, technology shock) by restricting the impact multiplier matrix. If we postulate a flexible price, sticky wage model with measurement error added to the real wage, for example, the static identifying restrictions are

$$\begin{pmatrix} u_t^{i} \\ u_t^{\text{real wage}} \\ u_t^{\text{inflation}} \\ u_t^{\text{output}} \\ u_t^{\text{hours}} \end{pmatrix} = \begin{bmatrix} + & + & + & - & * \\ - & + & - & + & * \\ + & - & + & - & * \\ - & - & + & + & * \\ - & - & + & - & * \end{bmatrix} \begin{pmatrix} w_t^{\text{markup}} \\ w_t^{\text{monetary}} \\ w_t^{\text{taste}} \\ w_t^{\text{technology}} \\ w_t^{\text{measurement error}} \end{pmatrix}.$$

Using simplified versions of this model, Canova and Paustian (2011) demonstrate that the ability of sign-identified VAR models to recover the

structural responses implied by DSGE models improves when more shocks are identified (even if those shocks are not the shocks of interest) and when more variables are restricted for a given number of shocks. Especially the estimation of responses to monetary policy shocks requires many restrictions. Canova and Paustian also provide evidence that sign-identified VAR models of lower dimension may still be capable of recovering the DSGE population responses, as long as the omitted shocks do not exhibit the same sign patterns as the shocks to be identified.

This result addresses to some extent concerns in Fry and Pagan (2011) that the approach of deriving sign restrictions from DSGE models relies on the DSGE model having the same reduced-form VAR representation that is imposed in applied work. They observe that in practice omitting some of the variables contained in the underlying DSGE model may undermine the ability of sign-identified VAR models to recover the population responses. Fry and Pagan's conclusion is that caution is called for in imposing sign restrictions derived from DSGE models, unless there is a one-to-one mapping between the VAR model and the DSGE model. The results in Canova and Paustian (2011) suggest that sign-identified VAR models are more robust than exactly identified VAR models to model misspecification such as omitted variables.

13.8.8 A Model of Speculation in the Global Oil Market

Kilian and Murphy (2014) and Kilian and Lee (2014) propose a monthly model of the global oil market that allows for an explicit role of speculators in the physical oil market. In addition to the percent change in global oil production ($\Delta prod$), a business cycle measure of global real economic activity (rea), and the log of the real price of oil ($rpoil$), this model includes the change in above-ground global crude oil inventories ($\Delta Inventories$). Given the seasonality of the inventory data, the model is estimated using seasonal dummies (see Chapter 19). With the inclusion of the inventory data, it is no longer possible to defend a recursively identified model. Instead, the structural shocks are identified based on a combination of sign restrictions and bounds on the short-run price elasticities of oil demand and oil supply.

The key identifying assumptions are restrictions on the signs of the impact responses of the four observables to the structural shocks. First, an unanticipated disruption in the flow supply of oil ($w_t^{\text{flow supply}}$) causes oil production to fall, the real price of oil to increase, and global real activity to fall on impact. Second, an unanticipated increase in the flow demand for oil ($w_t^{\text{flow demand}}$), defined as an increase in oil demand for current consumption, causes global oil production, global real activity, and the real price of oil to increase on impact. Third, a positive speculative demand shock ($w_t^{\text{speculative demand}}$) is defined as an increase in inventory demand driven by, for example, a higher expected

real price of oil that is not already captured by flow demand or flow supply shocks. Such a shock in equilibrium causes an accumulation of oil inventories and raises the real price of oil. The accumulation of inventories requires oil production to increase and oil consumption to fall (associated with a fall in global real activity). Finally, the model also includes a residual shock (w_t^{residual}) designed to capture idiosyncratic shocks driven by a myriad of reasons that cannot be classified as one of the first three structural shocks. This shock is defined implicitly as the complement to the remaining shocks. These assumptions imply that

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \\ u_t^{\Delta Inventories} \end{pmatrix} = \begin{bmatrix} - & + & + & * \\ - & + & - & * \\ + & + & + & * \\ * & * & + & * \end{bmatrix} \begin{pmatrix} w_t^{\text{flow supply}} \\ w_t^{\text{flow demand}} \\ w_t^{\text{speculative demand}} \\ w_t^{\text{residual}} \end{pmatrix}.$$

Note that the inventory responses to the flow supply and flow demand shocks are left unrestricted, but can be shown to be negative in the data, consistent with stabilizing inventory responses.

In addition to these static sign restrictions, the model imposes the additional restriction that the response of the real price of oil to a negative flow supply shock must be positive for at least twelve months, starting in the impact period. This restriction is necessary to rule out structural models in which unanticipated flow supply disruptions cause a decline in the real price of oil below its starting level. Such a decline would be at odds with conventional views of the effects of unanticipated oil supply disruptions. Because the positive response of the real price of oil tends to be accompanied by a persistently negative response of oil production, once we impose this additional dynamic sign restriction, it furthermore must be the case that global real activity responds negatively to oil supply shocks. This is the only way for the oil market to experience higher prices and lower quantities in practice, because in the data the decline of inventories triggered by an oil supply disruption is much smaller than the shortfall of oil production. This implies a joint set of dynamic sign restrictions such that the responses of oil production and global real activity to an unanticipated flow supply disruption are negative for the first twelve months, while the response of the real price of oil is positive.

Finally, the model imposes the restrictions that the impact price elasticity of oil supply is close to zero and that the impact price elasticity of oil demand cannot exceed the long-run price elasticity of oil demand, consistent with conventional views in the literature. A benchmark for that long-run elasticity is provided by studies of nonparametric gasoline demand functions based on U.S. household survey data such as Hausman and Newey (1995) which have consistently produced long-run price elasticity estimates near -0.8. Their estimate suggests a bound of -0.8 on the impact price elasticity of demand. The

construction of this price elasticity of oil demand is complicated by the presence of oil inventories. For details the reader is referred to Kilian and Murphy (2014).

The model's focus on above-ground crude oil inventories is consistent with conventional accounts of speculation involving the accumulation of oil inventories in oil-importing economies. An alternative view is that speculation may also be conducted by oil producers who have the option of leaving oil below the ground in anticipation of rising prices (see Hamilton 2009). An accumulation of below-ground inventories by oil producers in anticipation of rising prices is equivalent to a reduction in flow supply. In short, flow supply shocks and speculative supply shocks are observationally equivalent in this model.

13.9 Mixing Sign and Exclusion Restrictions

It is possible to combine sign restrictions with other identifying restrictions such as short-run or long-run exclusion restrictions. Of course, mixing sign and exclusion restrictions further complicates inference. Little is known at this point about how to provide valid summary statistics for such models. Empirical studies have tended to rely on the posterior median response functions and posterior quantile bands.

One situation encountered in applied work is a mixture of sign restrictions and long-run exclusion restrictions. As in the case of models mixing short-run and long-run exclusion restrictions, special care must be exercised in specifying the reduced-form model and in imposing the long-run restrictions when constructing the initial recursive model, for which alternative rotations are considered. Fisher, Huh, and Pagan (2016) provide an illustrative example, building on the analysis in Peersman (2005). A more common situation in applied work is the combination of sign restrictions with short-run exclusion restrictions.

13.9.1 Examples of Models Mixing Sign and Short-Run Zero Restrictions

A Model of Fiscal Policy Shocks. Mountford and Uhlig (2009) propose a quarterly structural VAR model for the U.S. economy including ten variables designed to identify a generic business cycle shock, a monetary policy shock, a government spending shock, and a government revenue shock. The model includes real GDP, real consumption, total government expenditures, total government revenue, real wages, private nonresidential investment, the short-term interest rate, adjusted reserves, the producer price index for crude materials, and the GDP deflator. All components of national income are in per capita terms. All variables but the interest rate are in logs.

A positive business cycle shock is defined as a shock that increases output, consumption, investment, and government revenue for the first year following the shock. A positive monetary policy shock increases interest rates. It also lowers adjusted reserves and all prices for the first year. The monetary policy shock is defined to be orthogonal to the business cycle shock. A positive government spending shock increases government spending for one year after the shock. A positive unanticipated government revenue shock, in contrast, increases government revenue for one year after a positive shock. The two fiscal policy shocks are defined to be orthogonal to the business cycle shock and the monetary policy shock, but not necessarily orthogonal with respect to each other. Mountford and Uhlig (2009) also consider an alternative identification allowing for anticipated government revenue shocks. In that case, government revenue is restricted to rise only one year after the shock, which implies exclusion restrictions on the earlier responses, followed by a positive response in the second year.

In this model, no explicit sign restrictions are imposed on the responses of output, investment, and consumption. Because the model is estimated using a penalty function along the lines discussed earlier, however, as shown in Arias, Rubio-Ramírez, and Waggoner (2013), there are implicit sign restrictions on the responses of these variables to fiscal policy shocks. If the same model is estimated imposing only the sign restrictions discussed by Mountford and Uhlig and not using the penalty function approach, there is no support for Mountford and Uhlig's main conclusion that deficit-financed tax cuts are better for stimulating economic activity than deficit-financed increases in government spending. In contrast, when the VAR model is estimated imposing additional sign restrictions on the responses of GDP, consumption, and investment, the credible set narrows and qualitatively the same results as in the penalty function approach emerge.

A Model of Shocks to Optimism about the Economy. In related work, Beaudry, Nam, and Wang (2012) consider a structural VAR model including a suitably adjusted measure of total factor productivity, the stock price, real consumption, the real federal funds rate, and hours worked. The baseline structural model only imposes the restriction that positive shocks to optimism about the state of the U.S. economy have no contemporaneous effect on adjusted total factor productivity, but raise stock prices. Beaudry et al. estimate this model using a penalty function as in Mountford and Uhlig (2009). They show that a positive shock to optimism triggers an increase in hours worked and consumption. If correct, this result would seem to endorse the view that business cycles are driven by spells of optimism and pessimism.

As Arias, Rubio-Ramírez, and Waggoner (2013) show, however, the use of a penalty function in solving this model implies the additional restriction that

positive shocks to optimism generate an increase in consumption and hours worked. Thus, it is not the case that Beaudry et al.'s conclusion only relies on the identifying assumption that positive shocks to optimism have no contemporaneous effect on productivity, but raise stock prices. Upon estimating the model without the penalty function, these results disappear and the estimates become uninformative. They reappear only if the additional sign restrictions are explicitly imposed in estimation.

A Small Open Economy Model. A third example, motivated by the analysis in Mumtaz and Surico (2009), is a small open economy model. All global variables are marked with an asterisk. The model includes global real GDP growth (Δgdp^*), global inflation (Δp^*), global money growth (Δm^*), and the global interest rate (i^*). It also includes real GDP growth (Δgdp), inflation (Δp), and the short-term interest rate (i) in the U.K. domestic economy.

The model has an international block and a domestic block. The international block is ordered first and consists of a goods market and a money market. A positive global aggregate demand shock raises all global macroeconomic aggregates on impact. A positive global supply shock raises global real output, but lowers global inflation on impact. The responses of global money growth and the global interest rate remain unrestricted. A positive shock to global money demand causes a decline in inflation and real output on impact as well as an increase in the global interest rate and money growth. A positive shock to global money supply causes a decline in the global interest rate on impact as well as an increase in real output, global inflation, and global monetary aggregates.

The domestic block consists of the last three equations. The domestic monetary policy shock is identified recursively as in standard semistructural models of monetary policy. The other two domestic shocks are not separately identified. A key identifying assumption is that there is no immediate feedback from the three domestic shocks to any of the global macroeconomic aggregates, resulting in a block of exclusion restrictions in the upper right corner of B_0^{-1}:

$$\begin{pmatrix} u_t^{gdp,*} \\ u_t^{\Delta p,*} \\ u_t^{\Delta m,*} \\ u_t^{i,*} \\ u_t^{\Delta p} \\ u_t^{\Delta gdp} \\ u_t^{i} \end{pmatrix} = \begin{bmatrix} + & + & - & + & 0 & 0 & 0 \\ + & - & - & + & 0 & 0 & 0 \\ + & * & + & + & 0 & 0 & 0 \\ + & * & + & - & 0 & 0 & 0 \\ * & * & * & * & * & 0 & 0 \\ * & * & * & * & * & * & 0 \\ * & * & * & * & * & * & * \end{bmatrix} \begin{pmatrix} w_t^{\text{AD},*} \\ w_t^{\text{AS},*} \\ w_t^{\text{money demand},*} \\ w_t^{\text{money supply},*} \\ w_{5t} \\ w_{6t} \\ w_t^{\text{monetary policy, UK}} \end{pmatrix}. \tag{13.9.1}$$

13.9.2 How to Combine Sign Restrictions and Exclusion Restrictions

A number of algorithms have been proposed for combining sign restrictions with selected short-run or long-run exclusion restrictions in special cases. Early examples include Baumeister and Peersman (2013), Baumeister and Benati (2013), and Benati and Lubik (2014). Of special interest is the use of subrotations to generate partially restricted draws of B_0^{-1} from block recursive models. A more general algorithm for combining sign restrictions, short-run exclusion restrictions, and long-run exclusion restrictions (including linear restrictions on the structural parameters themselves and on Q) has been proposed by Arias, Rubio-Ramírez, and Waggoner (2013). Closely related work includes Binning (2013).

Subrotations for Preserving Blocks of Zero Restrictions on B_0^{-1}**.** Block recursive models such as the small open economy example above are often solved using subrotations. Consider the small open economy example. Partition the K model variables into K_1 global variables and K_2 domestic variables. Note that there is no contemporaneous feedback from the K_2 domestic variables to the K_1 global variables in (13.9.1). The purpose of using subrotations is to preserve these exclusion restrictions in generating draws for the rotation matrix Q. The algorithm involves four steps.

Algorithm (*Subrotations*)

1. Start with the $K \times K$ lower-triangular Cholesky decomposition P such that $PP' = \Sigma_u$.
2. Generate a $K_1 \times K_1$ dimensional subrotation $\overline{Q}$ by drawing the columns of a $K_1 \times K_1$ matrix $\overline{W}$ from the $\mathcal{N}(0, I_{K_1})$ distribution and apply the QR decomposition $\overline{W} = \overline{Q}R$.
3. Form Q by placing $\overline{Q}$ in the upper left corner of the identity matrix I_K:
$$Q = \begin{bmatrix} \overline{Q} & 0 \\ 0 & I_{K_2} \end{bmatrix}.$$
4. Then a draw for B_0^{-1} consists of PQ. □

A numerical example for $K_1 = 3$ and $K_2 = 3$ illustrates this procedure. Let

$$P = \begin{bmatrix} 2.2295 & 0 & 0 & 0 & 0 & 0 \\ -2.3258 & 12.5955 & 0 & 0 & 0 & 0 \\ 0.1360 & 0.0746 & 0.8563 & 0 & 0 & 0 \\ -0.0383 & 0.1642 & 0.0222 & 0.3596 & 0 & 0 \\ -0.0019 & 0.0871 & 0.2364 & 0.1466 & 0.6272 & 0 \\ 0.0240 & 0.1288 & 0.4134 & -0.0344 & 0.5960 & 2.4835 \end{bmatrix}$$

and

$$\overline{W} = \begin{bmatrix} -0.7937 & 2.1457 & 0.0984 \\ -0.4165 & -0.7860 & -0.6739 \\ -0.7826 & 0.4481 & 0.2004 \end{bmatrix}.$$

Then

$$\overline{Q} = \begin{bmatrix} -0.6670 & 0.6467 & -0.3699 \\ -0.3500 & -0.7104 & -0.6106 \\ -0.6577 & -0.2778 & 0.7002 \end{bmatrix}.$$

This allows the construction of Q and hence

$$PQ = \begin{bmatrix} -1.4872 & 1.4418 & -0.8248 & 0 & 0 & 0 \\ -2.8575 & -10.4513 & -6.8309 & 0 & 0 & 0 \\ -0.6800 & -0.2030 & 0.5037 & 0 & 0 & 0 \\ -0.0465 & -0.1476 & -0.0705 & 0.3596 & 0 & 0 \\ -0.1847 & -0.1288 & 0.1130 & 0.1466 & 0.6272 & 0 \\ -0.3329 & -0.1908 & 0.2020 & -0.0344 & 0.5960 & 2.4835 \end{bmatrix}.$$

Partitioning the matrix $P = [P_1, P_2]$, where P_i is $K \times K_i$, $i = 1, 2$, we have $PQ = [P_1\overline{Q}, P_2]$. Thus, the elements of the $K_2 \times K_2$ block on the lower right of PQ remain unaffected by the rotations because they only depend on domestic coefficients that are exactly identified. The algorithm also preserves the $K_1 \times K_2$ block of zeros on the upper right that reflects the block recursive structure of the structural impact multiplier matrix. In contrast, the $K_2 \times K_1$ block on the lower left of PQ that relates to the global variables changes across rotations, as does the $K_1 \times K_1$ block on the upper left of PQ.

A More General Algorithm. One important limitation of the use of subrotations is that this approach relies on a block-recursive structure of the structural model. This restriction can be relaxed, as shown in Arias, Rubio-Ramírez, and Waggoner (2013). Express the stationary, K-dimensional structural VAR model as

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t. \tag{13.9.2}$$

The reduced-form representation of this model is

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $A_i = B_0^{-1} B_i$, $i = 1, \ldots, p$, $u_t = B_0^{-1} w_t$, and $\mathbb{E}(u_t u_t') = \Sigma_u = B_0^{-1} B_0^{-1\prime}$. Thus, the reduced form parameters are $A \equiv [A_1, \ldots, A_p]$ and Σ_u.

The impulse response function for the i^{th} variable with respect to the j^{th} structural shock at a finite horizon h corresponds to the element in row i and column j of

$$\Theta_h = (J \mathbf{A}^h J') B_0^{-1},$$

where $\mathbf{A}$ denotes the companion matrix of the reduced-form model,

$$\mathbf{A} \equiv \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_K & 0 & \cdots & 0 & 0 \\ 0 & I_K & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{bmatrix}$$

and $J \equiv [I_K, 0_{K \times K(p-1)}]$. When imposing restrictions on the infinite horizon, it is assumed that the i^{th} variable is expressed in first differences. The long-run structural impulse response function corresponds to the element in row i and column j of the matrix

$$\Theta_\infty \equiv \left(I_K - \sum_{i=1}^{p} A_i \right)^{-1} B_0^{-1}.$$

Candidate draws for Θ_∞ and Θ_h, denoted L_∞ and L_h, respectively, are constructed from the reduced-form estimates, given an initial guess of the structural impact multiplier matrix B_0^{-1} of $L_0 = \text{chol}(\widehat{\Sigma}_u)$. It is convenient to stack the structural impulse response matrices to be restricted into a single matrix denoted by $\mathbf{L}$. For example, if the sign restrictions are imposed at horizons zero, two, and infinity, then

$$\mathbf{L} = \begin{bmatrix} L_0 \\ L_2 \\ L_\infty \end{bmatrix}$$

has K columns corresponding to the K structural shocks and $3K$ rows because we consider restrictions at three horizons for K variables.

Sign restrictions on these structural impulse response functions can be represented by matrices S_j for $j = 1, \ldots, K$, where the number of columns in S_j equals the number of rows in $\mathbf{L}$. Usually S_j will be a selection matrix with one non-zero entry in each row. Let $s_j = \text{rank}(S_j)$. Then s_j is the number of sign restrictions on the impulse response functions associated with the j^{th} structural shock. The total number of sign restrictions is $s = \sum_{j=1}^{K} s_j$. Let ι_j denote the j^{th} column of the identity matrix I_K. Then the structural parameters satisfy the sign restrictions if and only if $S_j \mathbf{L} \iota_j > 0$ for $j = 1, \ldots, K$. As before, let Q denote the solution to the QR decomposition of a $K \times K$ matrix whose columns are independent random draws from $\mathcal{N}(0, I_K)$.

The challenge is to convert draws from the posterior of the reduced-form parameters, together with a draw for Q from the uniform distribution over the space of orthogonal matrices $\mathcal{O}(K)$, to a draw from the posterior distribution of the impulse responses before any identifying restrictions are imposed. The

following algorithm from Arias, Rubio-Ramírez, and Waggoner (2013) can be used for that purpose.

Algorithm (*Sign Restrictions*)

1. Draw (A, Σ_u) from the posterior distribution of the reduced-form parameters.
2. Draw an orthogonal matrix Q from $\mathcal{O}(K)$.
3. Then $L_0 Q$ will be a draw from the posterior distribution of B_0^{-1} before any identifying restrictions are imposed on the implied structural impulse responses.
4. Recompute $\mathbf{L}$ with $L_0 Q$ replacing L_0. Retain the draw if $S_j \mathbf{L} \iota_j > 0$ is satisfied for $j = 1, \ldots, K$.
5. Having repeated steps 2, 3, and 4 as often as desired, return to step 1 until the desired number of draws from the posterior of the structural impulse responses conditional on the sign restrictions has been obtained. □

This algorithm has to be modified to allow for additional exclusion restrictions on B_0^{-1}. Because the set of structural parameters conditional on zero restrictions has probability measure zero in the space of all structural parameters generated by the algorithm of Rubio-Ramírez, Waggoner, and Zha (2010), it is necessary to generate draws from the posterior of the structural impulse responses conditional on the zero restrictions already having been imposed, as outlined in Arias, Rubio-Ramírez, and Waggoner (2013).

Similar to the case of sign restrictions, we stack the structural impulse response functions at the horizons of interest. $\mathbf{L}$ now contains impulse response functions subject to both sign and exclusion restrictions. Zero restrictions are represented by matrices E_j for $j = 1, \ldots, K$. The number of columns in E_j is equal to the number of rows in $\mathbf{L}$. If the rank of E_j is e_j, then e_j is the number of zero restrictions associated with the j^{th} shock. The total number of zero restrictions is $e = \sum_{j=1}^{K} e_j$. The structural parameters satisfy the zero restrictions if and only if $E_j \mathbf{L} \iota_j = 0$ for $1 \leq j \leq K$. In what follows, let E_j represent the zero restrictions with the equations of model (13.9.2) ordered such that $e_j \leq K - j$ for $j = 1, \ldots, K$.

Algorithm (*Combining Sign Restrictions and Exclusion Restrictions*)

1. Draw (A, Σ_u) from the posterior distribution of the reduced-form parameters.
2. Draw an orthogonal matrix Q such that $L_0 Q$ satisfies the exclusion restrictions.
3. Recompute $\mathbf{L}$ with $L_0 Q$ replacing L_0. Retain the draw if $S_j \mathbf{L} \iota_j > 0$ is satisfied for $j = 1, \ldots, K$.

4. Having repeated steps 2 and 3 as often as is desired, return to step 1 until the desired number of draws from the posterior of the structural impulse responses conditional on the sign restrictions has been obtained. □

The modification in the algorithm involves step 2. Let X be a $K \times K$ matrix of independent $\mathcal{N}(0, 1)$ draws and choose Q by a QR decomposition of X. Define $q_j \equiv Q\iota_j$ and $x_j \equiv X\iota_j$ for $1 \leq j \leq K$, where ι_j denotes the j^{th} column of the identity matrix I_K. Moreover, define $R_1 = E_1\mathbf{L}$ and, for $j = 2, \ldots, K$,

$$R_j = \begin{bmatrix} E_j\mathbf{L} \\ Q_{j-1} \end{bmatrix},$$

where $\mathbf{L}$ denotes the matrix of structural impulse responses to be restricted, E_j is the corresponding matrix of zero restrictions, and $Q_{j-1} = [q_1, \ldots, q_{j-1}]$. Then drawing Q involves the following recursive subroutine:

a. Let $j = 1$.
b. Construct R_j and find the matrix N_{j-1} whose columns form an orthonormal basis for the null space of R_j.[6]
c. Draw the $K \times 1$ vector x_j from the $\mathcal{N}(0, I_K)$ distribution.
d. Let $q_j = N_{j-1}\left(N'_{j-1}x_j / \left\|N'_{j-1}x_j\right\|\right)$, where $\|\cdot\|$ is the Euclidian norm.[7]
e. If $j = K$, stop, construct $Q = [q_1, \ldots, q_K]$ and move to step 3 of the modified algorithm. Otherwise, let $j = j + 1$ and move to step b.

As Arias, Rubio-Ramírez, and Waggoner (2013) emphasize, when implementing this modified algorithm, it is critical that one exits upon finding a q_i that violates the sign restrictions, and resumes with a new draw from the reduced-form posterior in step 1. Although it is not permissible to draw orthogonal matrices until acceptance, it is permissible to draw a fixed number of orthogonal matrices Q for each reduced-form draw and to retain all rotations that satisfy the sign restrictions.

A Numerical Example. The following numerical VAR example from Arias, Rubio-Ramírez, and Waggoner (2013) illustrates the implementation of the modified algorithm. Let $K = 4$ and $p = 1$. Let A and Σ_u be a particular draw from the posterior of the reduced-form parameters of a four-dimensional

[6] An orthonormal basis for the null space of the matrix R_j may be obtained from a singular value decomposition applying the MATLAB function *null* to R_j, for example.

[7] The Euclidian norm of a vector $x = (x_1, \ldots, x_n)'$ is $\sqrt{x'x}$ which can be computed using the *norm* function in MATLAB.

VAR(1):

$$A_1 = \begin{bmatrix} 0.7577 & 0.7060 & 0.8325 & 0.4387 \\ 0.7431 & 0.0318 & 0.6948 & 0.3816 \\ 0.3922 & 0.2769 & 0.3171 & 0.7655 \\ 0.6555 & 0.0462 & 0.9502 & 0.7952 \end{bmatrix},$$

$$\Sigma_u = \begin{bmatrix} 0.0281 & -0.0295 & 0.0029 & 0.0029 \\ -0.0295 & 3.1850 & 0.0325 & -0.0105 \\ 0.0029 & 0.0325 & 0.0067 & 0.0054 \\ 0.0029 & -0.0105 & 0.0054 & 0.1471 \end{bmatrix}.$$

Suppose that we want to impose restrictions on the structural impulse response functions at horizons 0, 2, and ∞. First, without loss of generality, define an initial solution for the structural impulse response functions of interest from the reduced-form parameters as

$$\begin{aligned} L_0 &= \text{chol}(\widehat{\Sigma}_u), \\ L_2 &= (J\mathbf{A}^2J')L_0 = A_1^2 L_0, \\ L_\infty &= (I_4 - A_1)^{-1} L_0, \end{aligned}$$

where $\mathbf{A} = A_1$ because we are considering a VAR(1). Hence,

$$\mathbf{L} = \begin{bmatrix} L_0 \\ L_2 \\ L_\infty \end{bmatrix} = \begin{bmatrix} 0.1676 & 0 & 0 & 0 \\ -0.1760 & 1.7760 & 0 & 0 \\ 0.0173 & 0.0200 & 0.0775 & 0 \\ 0.0173 & -0.0042 & 0.0669 & 0.3772 \\ 0.1355 & 1.9867 & 0.1828 & 0.5375 \\ 0.0259 & 1.3115 & 0.0828 & 0.2882 \\ 0.1377 & 2.1813 & 0.2131 & 0.6144 \\ 0.1069 & 2.0996 & 0.1989 & 0.6281 \\ 0.1091 & -0.3783 & -0.0847 & -0.2523 \\ -0.1170 & 1.2928 & -0.0599 & -0.2201 \\ -0.0422 & -0.7342 & 0.0006 & -0.1695 \\ -0.0575 & -1.1662 & 0.0362 & 0.2577 \end{bmatrix}.$$

The identifying restrictions to be imposed on these structural responses are as follows. The response of the third variable to the second structural shock at horizon 2 is negative; the response of the fourth variable to the second structural shock is positive at horizon 2. The response of the second variable to the third structural shock is negative at horizon 0; and the response of the first variable to the fourth structural shock is positive at horizons 0, 2, and ∞. The responses of the first and third variables to the first shock are zero at horizon

0; and the response of the fourth variable to the second structural shock is zero at horizon ∞. These restrictions imply that

$$S_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$S_3 = \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$S_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$E_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

There are no sign restrictions associated with the first structural shock, so we do not need to specify S_1, and there are no exclusion restrictions associated with the third and fourth structural shock, so there is no need to specify E_3 and E_4.

To find a rotation matrix Q that satisfies both the sign and zero restrictions, we apply the subroutine described above:

a. Let $j = 1$.

b. Then $R_1 = \begin{bmatrix} 0.1676 & 0 & 0 & 0 \\ 0.0173 & 0.0200 & 0.0775 & 0 \end{bmatrix}$ and $N_0 = \begin{bmatrix} 0 & 0 \\ -0.0982 & 0 \\ 0.2502 & 0 \\ 0 & 1 \end{bmatrix}$.

c. Suppose the draw of the vector x_1 from the $\mathcal{N}(0, I_4)$ distribution is

$$x_1 = \begin{bmatrix} 0.4395 & -0.1190 & -0.9354 & 0.0464 \end{bmatrix}'.$$

d. Then $q_1 = N_0 \left(N_0' x_1 / \left\| N_0' x_1 \right\| \right) = \begin{bmatrix} 0 & 0.9018 & -0.2330 & 0.3638 \end{bmatrix}'$.

e. Let $j = 2$ and move to step b.

By repeating this routine for $j = 2, \ldots, 4$, given the random draws

$$x_2 = \begin{bmatrix} -0.6711 \\ 1.5332 \\ -0.1836 \\ 0.3509 \end{bmatrix}, \quad x_3 = \begin{bmatrix} -0.5941 \\ 0.5901 \\ -1.4499 \\ -0.2632 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 0.6713 \\ -0.4112 \\ 0.7989 \\ -0.0868 \end{bmatrix},$$

we obtain the following solutions:

$$N_1 = \begin{bmatrix} -0.2648 & 0.9609 \\ 0.1095 & -0.0053 \\ 0.9068 & 0.2271 \\ 0.3093 & 0.1586 \end{bmatrix}, \; N_2 = \begin{bmatrix} 0.1704 & -0.0322 \\ 0.2180 & -0.3697 \\ 0.9582 & 0.0188 \\ 0.0733 & 0.9284 \end{bmatrix},$$

$$N_3 = \begin{bmatrix} -0.0854 \\ -0.4203 \\ -0.2913 \\ 0.8551 \end{bmatrix}.$$

Upon completing this subroutine, we obtain the following solution for the restricted rotation matrix:

$$Q = [q_1, \ldots, q_K] = \begin{bmatrix} 0 & -0.9849 & -0.1509 & 0.0854 \\ 0.9018 & 0.0498 & -0.0871 & 0.4203 \\ -0.2330 & 0.1651 & -0.9130 & 0.2913 \\ 0.3638 & -0.0177 & -0.3189 & -0.8551 \end{bmatrix}.$$

The implied posterior draw of the structural impact multiplier matrix is

$$L_0 Q = \begin{bmatrix} 0 & -0.1651 & -0.0253 & 0.0143 \\ 1.6016 & 0.2617 & -0.1281 & 0.7313 \\ 0 & -0.0033 & -0.0751 & 0.0325 \\ 0.1179 & -0.0129 & -0.2025 & -0.3034 \end{bmatrix}.$$

The restrictions may be verified using $S_j \mathbf{L} q_j$ and $E_j \mathbf{L} q_j$, where $\mathbf{L}$ is evaluated at $L_0 Q$.

Caveats. Although the algorithm of Arias, Rubio-Ramírez, and Waggoner (2013) represents an important step forward in implementing mixed restrictions, several open questions remain. One limitation of this algorithm is that we can impose at most $K - j$ zero restrictions for each shock j, where $j = 1, \ldots, K$. In addition, this approach cannot be directly extended to imposing nonzero restrictions on specific elements of B_0. In contrast, additional inequality restrictions on structural impulse responses could easily be imposed in step 3 by generating draws for the structural impulse responses in question and dropping draws that do not satisfy these additional restrictions. One example of such additional restrictions would be the restriction that the impact response of one variable should be larger than the impact response of another variable to the same structural shock, as in Peersman (2005). Another example are bounds on impact elasticity parameters as in Kilian and Murphy (2014). A third example is the use of shape restrictions as in Scholl and Uhlig (2008).

A second limitation relates to the evaluation of the set of admissible models. Arias, Rubio-Ramírez, and Waggoner (2013) rely on posterior median

response functions and quantile bands. This approach suffers from the shortcomings already discussed earlier. It therefore would be useful to generalize the approach of constructing the most likely structural model, as discussed in Section 13.6.2, to the framework of Arias et al. and to derive the analogous expression for the posterior density of the structural impulse responses. Such an extension of the analysis in Inoue and Kilian (2013) should be straightforward in principle.

13.9.3 Discussion

As in the case of models identified only by sign restrictions, the modified algorithms discussed in this section will in general be informative for the structural impulse responses. Four proposals have been made to address this concern. One proposal has been to adapt the procedure of Giacomini and Kitagawa (2015) that is robust to the choice of the rotation matrix.

An alternative proposal has been to focus on priors that are conditionally agnostic. Arias, Rubio-Ramírez, and Waggoner (2015) observe that if a researcher begins with an agnostic prior over a particular parameterization of the model, as defined earlier, and then conditions on the zero restrictions, the resulting prior will be conditionally agnostic over that parameterization of the model. Unlike in the case of imposing sign restrictions only, however, one has to choose which parameterization of the model one wishes to be conditionally agnostic about. Being conditionally agnostic under one parameterization of the structural model does not imply being conditionally agnostic under a different parameterization. Arias et al. propose using an importance sampler to draw from the conditionally agnostic prior for a given parameterization of the structural VAR model. They also show how to construct conditionally flat priors by choosing an appropriate conjugate prior over the orthogonal reduced-form representation, generalizing the approach discussed earlier for the case of sign restrictions only. The choice of this conjugate prior depends on whether one wishes to be flat across the structural impulse responses that satisfy the zero restrictions or flat across the structural model representations that satisfy the zero restrictions.

Yet another proposal has been to embrace the fact that priors for structural impulse responses are informative and to make that information explicit, possibly drawing on extraneous economic information. Plagborg-Møller (2016) suggests such an algorithm that facilitates the imposition of zero restrictions on elements of B_0^{-1}. A complementary proposal has been to specify the prior on the elements of B_0 rather than B_0^{-1}, taking the structural VAR model representation as the object of primary interest (see Baumeister and Hamilton 2015a, 2015b). This procedure also allows us to interpret exclusion restrictions as degenerate priors.

13.10 Empirical Illustrations

We conclude this chapter with two empirical illustrations of the use of sign-identified VAR models.

13.10.1 A Model of the Global Oil Market

The first example is based on the analysis in Inoue and Kilian (2013). Inoue and Kilian adapt the monthly recursively identified structural VAR model of the global oil market in Kilian (2009) for the use of sign restrictions, building on Kilian and Murphy (2012). A negative flow supply shock initially lowers global oil production and global real economic activity, but raises the real price of oil. A positive flow demand shock initially raises oil production, real economic activity, and the real price of oil. Finally, other oil demand shocks (such as precautionary demand shocks) stimulate global oil production and the real price of oil, but lower real economic activity on impact. Formally:

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} - & + & + \\ - & + & - \\ + & + & + \end{bmatrix} \begin{pmatrix} w_t^{\text{flow supply}} \\ w_t^{\text{flow demand}} \\ w_t^{\text{other demand}} \end{pmatrix}.$$

In addition, the model numerically bounds the price elasticity of oil supply by 0.025, building on Kilian and Murphy (2012).

This elasticity corresponds to the ratio of the oil production response in the impact period triggered by an exogenous demand shock to the price response in the impact period triggered by the same shock. This restriction rules out models with implausibly high price elasticities of oil supply compared with conventional wisdom and extraneous empirical evidence. Finally, Inoue and Kilian (2013) restrict the real price of oil to be positive for the first year in response to positive demand and negative supply shocks, following Baumeister and Peersman (2013).

Following Inoue and Kilian (2013) and related studies in the literature, we specify a VAR(24) model with intercept. The model is estimated on monthly data for 1973m2-2008m9 using a diffuse Gaussian-inverse Wishart prior. Figure 13.7 plots the structural responses. The responses have been normalized such that each structural shock implies an increase in the real price of oil. The response of oil production is obtained by cumulating the responses of its growth rate.

All structural response function estimates are consistent with standard economic intuition. For example, a negative flow supply shock is associated with a persistent decline in oil production, a modest increase in the real price of oil, and a short-lived decline in global real economic activity. A positive flow demand shock is associated with a persistent and hump-shaped response

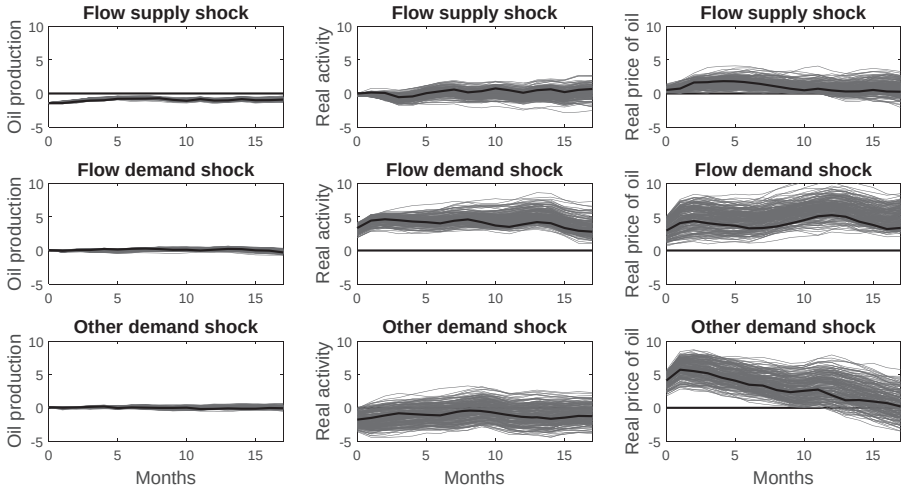

Figure 13.7. Sign-identified oil market model impulse response functions in the modal model and 68% joint HPD regions.

in both global real activity and the real price of oil and with little response in global crude oil production. Other demand shocks (such as shocks to oil inventory demand) cause a temporary increase in the real price of oil, a persistent decline in global real economic activity and little response in global crude oil production. The corresponding credible sets indicate considerable uncertainty about the price responses and to a lesser extent for the responses in real activity, whereas the credible sets for oil production responses are quite narrow. Nevertheless, several response functions are precisely enough estimated to conclude that the response differs from zero. Figure 13.7 also illustrates that the responses of the most likely model need not be near the center of the credible set.

There are also important differences between the most likely estimates provided by the modal model and the conventional median response functions. Median response functions may be closer to zero or further away from zero than the responses of the modal model. For example, median response functions can be shown to overestimate the magnitude of the price response to flow demand shocks, as illustrated in the left panel of Figure 13.8. Moreover, pointwise 68% posterior error bands provide little protection against mischaracterizing the impulse response dynamics, as shown in the right panel of Figure 13.8. At several horizons, the response functions of the modal model are outside the pointwise error bands. The right panel of Figure 13.8 also illustrates that pointwise intervals tend to misrepresent the estimation and identification uncertainty compared with the credible set shown in Figure 13.7 that captures the joint uncertainty over all impulse responses. This example illustrates that

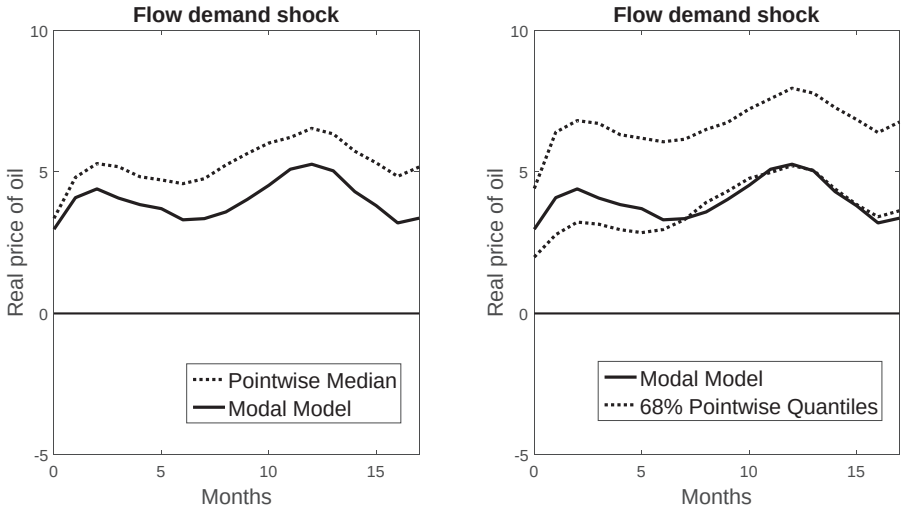

Figure 13.8. Structural impulse responses in the sign-identified oil market model.

the way estimates of sign-identified VAR models are represented matters for the interpretation of the estimated model.

13.10.2 A Model of Monetary Policy

Whereas the first example involved a fully identified model based on sign restrictions, the second example based on Uhlig (2005) involves a sign-identified model that is only partially identified. The set of variables consists of monthly U.S. data for the log of interpolated real GDP (gdp_t) and of its deflator ($defl_t$), the log of a commodity price index ($pcom_t$), total reserves (tr_t), nonborrowed reserves (nbr_t), and the federal funds rate (i_t). The identification is deliberately agnostic. An unanticipated monetary policy tightening is assumed to cause an increase in the interest rate and to lower the GDP price deflator, the commodity price index and nonborrowed reserves all for the first six months, including the impact period:

$$\begin{pmatrix} u_t^{gdp} \\ u_t^{defl} \\ u_t^{pcom} \\ u_t^{tr} \\ u_t^{nbr} \\ u_t^{i} \end{pmatrix} = \begin{bmatrix} * & * & * & * & * & * \\ - & * & * & * & * & * \\ - & * & * & * & * & * \\ * & * & * & * & * & * \\ - & * & * & * & * & * \\ + & * & * & * & * & * \end{bmatrix} \begin{pmatrix} w_t^{\text{monetary policy}} \\ w_{2t} \\ w_{3t} \\ w_{4t} \\ w_{5t} \\ w_{6t} \end{pmatrix},$$

where * denotes unrestricted elements.

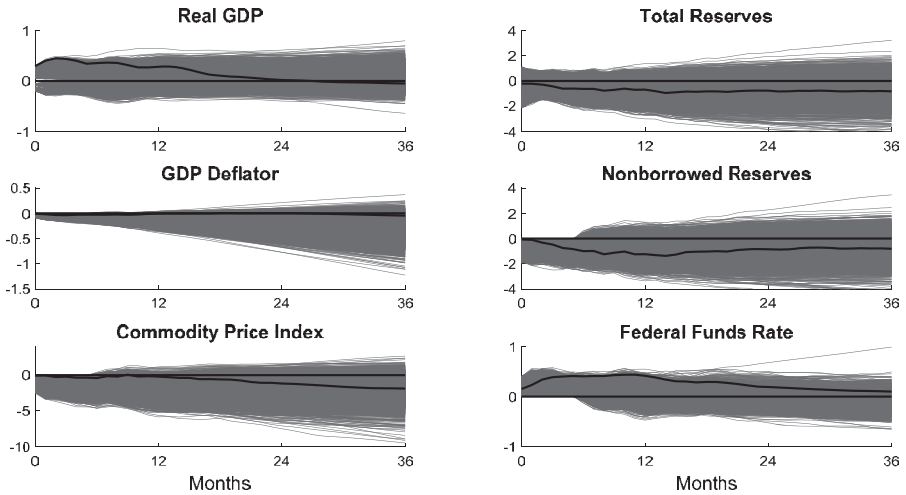

Figure 13.9. Responses to a monetary policy tightening in the original sign-identified model: Response functions in the modal model and 68% joint HPD regions.
Source: Inoue and Kilian (2013).

The sample period is 1965m1-2003m12 to ensure compatibility with Uhlig's original analysis. The VAR(12) model without intercept is estimated using a diffuse Gaussian-inverse Wishart prior on the same data as in Uhlig (2005). We evaluate the posterior of the model as in Inoue and Kilian (2013). The numerical stability of the results requires a fairly large number of draws, especially for M and $M^{\dagger}$, where $M^{\dagger}$ is the number of draws used to approximate the marginalized posterior distribution. The posterior distribution of the impulse response estimates is based on $M = 5{,}000$ draws from the reduced-form posterior distribution with $N = 500$ rotations each. We set $M^{\dagger} = 20{,}000$. Figure 13.9 summarizes the responses of the variables of interest to an unexpected monetary policy tightening.

Figure 13.9 illustrates that there are important differences between the median response estimates of the response of real output and the response in the modal model. Whereas Uhlig reports a peak median output response of 0.15 percentage points, for the same data, the modal model implies a peak response of almost 0.5 percentage points. Moreover, that peak value is near the upper end of the credible set and outside the conventional pointwise posterior error band. It should be noted that both the median estimate and the response estimate based on the modal model are counterintuitive in that a monetary tightening would be expected to cause a decline in real output over time rather than an increase. This outcome reflects the fact that the identifying assumptions are not overly informative. Even in Uhlig's original analysis, there was substantial pointwise probability mass on both negative and positive responses

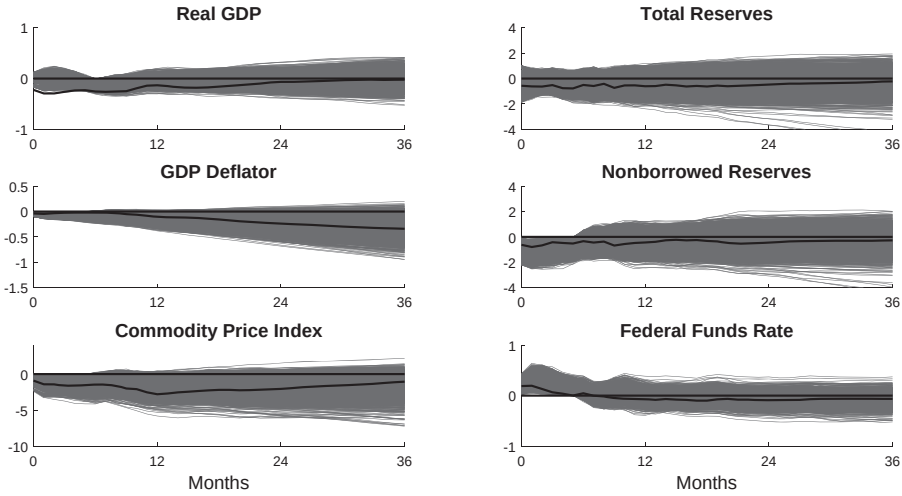

Figure 13.10. Responses to a monetary policy tightening in the modified sign-identified model: Response functions in the modal model and 68% joint regions of highest posterior density.
Source: Inoue and Kilian (2013).

of real output. The 68% credible set further widens the set of probable response functions.

The explicit reason why Uhlig (2005) did not impose further restrictions is that he wished to be as agnostic as possible about the response of real output. This approach is appropriate only to the extent that we view models in which real output increases in response to a monetary tightening as economically plausible a priori (see Kilian and Murphy 2012). Many economists would disagree with this view at least for intermediate horizons. Hence, in Figure 13.10 we consider an alternative set of results for models that impose an additional sign restriction on the response of real GDP to an unexpected monetary policy tightening after 6 months (and only at that horizon):

$$\frac{\partial gdp_{t+6}}{\partial w_t^{\text{monetary policy}}} < 0.$$

This identifying assumption leaves the short-run as well as the longer-run response of real output unrestricted, preserving the spirit of Uhlig's original exercise.

The resulting modal model produces substantially different and more economically plausible results, including a cumulative drop in real GDP of -0.3 percentage points in the second quarter. The response estimate for the modal model is at the lower end of the credible set and again outside the conventional pointwise posterior error band. It also is substantially different from the response estimate obtained from the traditional Cholesky decomposition (see

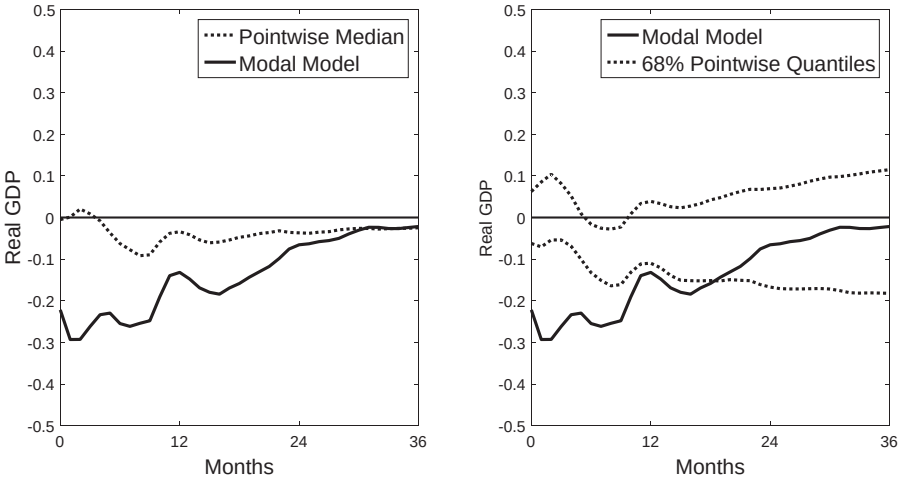

Figure 13.11. Responses to a monetary policy tightening in the modified sign-identified model.
Source: Inoue and Kilian (2013).

Chapter 12). One difference is that the reduction in real GDP in Figure 13.10 is temporary, whereas traditional Cholesky models imply a much more persistent decline in real GDP. Even in this alternative model, however, the 68% credible set includes many positive real output responses, suggesting that the data are not informative about the response of real output. Likewise the other response functions are estimated only very imprecisely. We conclude that there remains substantial uncertainty about the effects of monetary policy shocks on real output.

Figure 13.11 elaborates further on the results in Figure 13.10. The left panel again illustrates that there can be substantial differences between the median response function estimates and the response function estimates based on the modal model. For example, the decline in real GDP caused by an unanticipated monetary contraction is much larger in the modal model, at least in the short-run. In some cases, the median and the modal response of real GDP differ not only in magnitude but in sign. The right panel demonstrates that the modal model responses may be outside the conventional pointwise 68% error bands. This is true in particular for the response of real GDP and to a lesser extent for the response of commodity prices and the own-response of the federal funds rate.

13.11 Concluding Remarks

The advantage of sign-identified models is that sign restrictions are much more easily derived from economic models than alternative identifying restrictions.

At the same time, working with set-identified models poses special challenges. Of particular concern is the possibility that the conventional priors used in estimating sign-identified models may unduly influence the posterior of the structural impulse responses.

Making explicit prior information about the parameters of the structural VAR representation, as proposed by Baumeister and Hamilton (2015a, 2015b, 2015c) is one possible response to this concern. This approach also provides an opportunity for the researcher to bring to bear additional information that otherwise could not have been imposed. However, it is not always clear how to derive within this framework economically motivated priors that have broad appeal. Moreover, this approach typically does not allow the user to replicate the identifying restrictions used in other studies. For example, the methodology proposed by Baumeister and Hamilton (2015a) does not allow for dynamic sign restrictions, shape restrictions, and cross-equation restrictions. Even more importantly, the posterior draws generated by the approach of Baumeister and Hamilton (2015a) currently can only be evaluated under restrictive loss functions that postulate that the user is not concerned with the shapes of the impulse response functions. One way of addressing the latter problem would be to adapt the methodology of Inoue and Kilian (2013) to this new setting, but no such results exist at this point.

Another possible solution to the problem of unintentionally informative priors in sign-identified structural VAR models has been proposed by Plagborg-Møller (2016). His proposal is to estimate the structural moving average coefficients directly by Bayesian methods, allowing one to explicitly specify the prior densities for the structural impulse response parameters not only on impact but also at longer horizons. This approach avoids some of the drawbacks of the proposal in Baumeister and Hamilton (2015a). It not only follows the existing literature on sign-identified VAR models in specifying the prior on the structural impact multiplier matrix, but it also facilitates the imposition of restrictions on the shape and smoothness of the structural impulse response functions. In this sense Plagborg-Møller's approach may be viewed as the complement to Baumeister and Hamilton's proposal to impose priors on the parameters of the structural VAR model. In both cases, the central idea is to replace implicitly informative priors by explicitly informative priors. The main drawback of Plagborg-Møller's approach is that the resulting structural impulse response estimates are likely to be sensitive to the choice of the prior in finite samples and asymptotically. Moreover, as in Baumeister and Hamilton's work, it is rarely clear what independent information the specification of the prior density can be based on.

An alternative response to the concern that conventional priors may be unintentionally informative about the structural impulse responses is the use of the robust Bayesian approach of Giacomini and Kitagawa (2015), which provides bounds on the structural impulse responses. Those bounds, however, may be wide and uninformative.

A third response is to specify the prior to be agnostic (or conditionally agnostic) in the technical sense defined by Arias, Rubio-Ramírez, and Waggoner (2015) or to modify existing priors to imply a flat prior distribution over either the structural VAR representation or the structural impulse response representation. As in Baumeister and Hamilton (2015a), the posterior draws generated by these alternative algorithms to date have only been evaluated under restrictive loss functions.

A fourth response is to dispense with all priors and to rely on frequentist confidence sets for sign-identified structural impulse responses, as discussed in Moon, Schorfheide, and Granziera (2013), Gafarov, Meier, and Montiel Olea (2015a, 2015b), and Kitagawa, Montiel Olea, and Payne (2015), but the latter approach also tends to produce bounds that are too wide to be informative in practice.