8 Identification by Short-Run Restrictions

8.1 Introduction

Consider a K-dimensional time series y_t, $t = 1, \ldots, T$. We postulate that the DGP for y_t is well-approximated by a vector autoregression of finite order p. Our objective is to learn about the parameters of the structural vector autoregressive model

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t, \tag{8.1.1}$$

where w_t denotes a mean zero serially uncorrelated error term, also referred to as a structural innovation or structural shock. The error term is assumed to be unconditionally homoskedastic, unless noted otherwise. The nonsingular matrix B_0 governs the contemporaneous interaction between the model variables. All deterministic regressors have been suppressed for notational convenience. Equivalently the model can be written more compactly as

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

where $B(L) \equiv B_0 - B_1 L - B_2 L^2 - \cdots - B_p L^p$ is the autoregressive lag polynomial. The variance-covariance matrix of the structural error term is typically normalized such that

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

This means, first, that there are as many structural shocks as variables in the model. Second, structural shocks by definition are mutually uncorrelated, which implies that Σ_w is diagonal. Third, we normalize the variance of all structural shocks to unity. The latter normalization does not involve a loss of generality as long as the diagonal elements of B_0 remain unrestricted. We defer a discussion of alternative normalizations until the end of this section.

For model (8.1.1) to be considered a structural VAR model it is not sufficient for the elements of w_t to be uncorrelated. In addition, these shocks must be economically interpretable. It is worth noting that, in general, structural shocks do not correspond to particular model variables and, hence, have no natural unit of measurement. For example, in a VAR system consisting of only price and quantity, both the demand shock and the supply shock shift the price as well as the quantity. In fact, if price and quantity variables were mechanically associated with price and quantity shocks, this would be an indication that the proposed model is not truly structural.

Next we derive the reduced-form representation of this structural VAR model. This involves expressing y_t as a function of lags of y_t only. After pre-multiplying both sides of the structural VAR representation by B_0^{-1},

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

the model can be represented in reduced form as

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

where $A_i = B_0^{-1}B_i$, $i = 1, \ldots, p$, and $u_t = B_0^{-1}w_t$. The reduced-form innovations u_t are a weighted average of the structural shocks w_t. Equivalently this model can be written more compactly as

$$A(L)y_t = u_t,$$

where $A(L) \equiv I_K - A_1L - A_2L^2 - \cdots - A_pL^p$ denotes the autoregressive lag polynomial. Standard estimation methods allow us to obtain consistent estimates of the reduced-form parameters A_i, $i = 1, \ldots, p$, the reduced-form innovations u_t, and their covariance matrix $\mathbb{E}(u_t u_t') \equiv \Sigma_u$ (see Chapter 2).

Having estimated the reduced-form model, the question arises of how to recover the structural representation of the VAR model. Knowledge of B_0 enables us to reconstruct w_t from $w_t = B_0 u_t$ and B_i, $i = 1, \ldots, p$, from $B_i = B_0 A_i$. The central question therefore is how to recover the elements of B_0 (or its inverse) from consistent estimates of the reduced-form parameters.

By construction, $u_t = B_0^{-1}w_t$. Hence, the variance of u_t is

$$\Sigma_u = \mathbb{E}(u_t u_t') = B_0^{-1}\mathbb{E}(w_t w_t')B_0^{-1\prime} = B_0^{-1}\Sigma_w B_0^{-1\prime} = B_0^{-1}B_0^{-1\prime}, \tag{8.1.2}$$

where we made use of $\Sigma_w = I_K$. We can think of $\Sigma_u = B_0^{-1}B_0^{-1\prime}$ as a system of nonlinear equations in the unknown parameters of B_0^{-1}. Observe that due to the symmetry of Σ_u, (8.1.2) represents a system of $K(K+1)/2$ independent equations only. The reduced-form error covariance matrix Σ_u can be estimated consistently and hence is treated as known for the time being. The system of nonlinear equations (8.1.2) can be solved for the unknown parameters in B_0^{-1} using numerical methods, provided the number of unknown parameters in B_0^{-1} does not exceed the number of independent equations in (8.1.2). This involves imposing additional restrictions on selected elements of B_0^{-1}. Such restrictions may take the form of exclusion restrictions, proportionality restrictions, or other equality restrictions. The most common approach is to impose exclusion restrictions on selected elements of B_0^{-1} by forcing these elements to be zero.

To verify that all of the elements of the unknown matrix B_0^{-1} are uniquely identified, observe that Σ_u has $K(K+1)/2$ free parameters. This follows from the fact that any covariance matrix is symmetric about the main diagonal. Hence, $K(K+1)/2$ by construction is the maximum number of parameters in B_0^{-1} that one can uniquely identify. This order condition for just identification is easily checked in practice by counting the unrestricted elements of B_0^{-1}. It is a necessary condition for identification only, however.

Even if the order condition is satisfied, the system of equations may fail to have a unique solution. To ensure identification of the structural shocks, the system, in addition, has to satisfy the rank condition for identification. Rubio-Ramírez, Waggoner, and Zha (2010) discuss a general approach to evaluating the rank condition for global identification in structural VAR models. For further discussion of the identification issue, see also Christiano, Eichenbaum, and Evans (1999) and Taylor (2004).

It should be noted that even if all elements of w_t are uniquely identified in a statistical sense, they need not be uniquely identified in the economic sense. For example, there are structural VAR models in which all elements of w_t are statistically uniquely identified, but only some of the elements of w_t are also economically identified.

The earlier discussion alluded to the existence of alternative normalization assumptions in structural VAR analysis. There are three equivalent representations of structural VAR models that differ only in how the model is normalized. All three representations have been used in applied work. In the discussion thus far we made the standard normalizing assumption that $\Sigma_w = I_K$, while leaving the diagonal elements of B_0 unrestricted. Identification was achieved by imposing identifying restrictions on B_0^{-1}. By construction, a unit innovation in the structural shocks in this representation is an innovation of a magnitude of one standard deviation, so structural impulse responses are responses to one-standard deviation shocks.

Equivalently, one could have left the diagonal elements of Σ_w unconstrained and set the diagonal elements of B_0 to unity in $w_t = B_0 u_t$ (see, e.g., Keating 1992). However, the variance of the structural errors will no longer be unity under these assumptions, so the implied responses to shocks must be rescaled by one standard deviation of the structural residual to ensure that the implied structural impulse responses represent responses to one-standard deviation shocks. Under this alternative normalization, the order condition for just identification requires B_0 to have $K(K-1)/2$ restrictions, not counting the restrictions on the diagonal elements. Which convention one uses is usually dictated by the nature of the economic identifying assumptions. In some cases it is more natural to restrict B_0^{-1} and in others B_0.

Finally, these two approaches may be combined by changing notation and writing the model equivalently as

$$B_0 u_t = C w_t$$

with $\Sigma_w = I_K$ such that $\Sigma_u = B_0^{-1} C C' B_0^{-1\prime}$. The two representations above emerge as special cases of this representation with the alternative normalizations of $B_0 = I_K$ or $C = I_K$. The advantage of this third representation is that it allows one to relax the assumption that either $C = I_K$ or $B_0 = I_K$, which sometimes facilitates the exposition of the identifying assumptions.[1] If the diagonal elements of B_0 are normalized to unity and $\Sigma_w = I_K$, the order condition requires $K^2 + K(K-1)/2$ restrictions on B_0 and C combined, including the restrictions on the diagonal elements of B_0. Likewise, if the diagonal elements of both B_0 and C are restricted to unity instead and the diagonal elements of Σ_w are left unrestricted, the order condition calls for $K^2 + K(K+1)/2$ restrictions, including the normalizing restrictions on the diagonals. For example, Blanchard and Perotti (2002) use this representation with the diagonal elements of both B_0 and C normalized to unity, but neither C nor B_0 diagonal.

8.2 Recursively Identified Models

One popular way of disentangling the structural innovations w_t from the reduced-form innovations u_t is to "orthogonalize" the reduced-form errors. Orthogonalization here means making the errors mutually uncorrelated. Mechanically, this can be accomplished as follows. Define the lower-triangular $K \times K$ matrix P with positive main diagonal such that $PP' = \Sigma_u$. The matrix P is known as the lower-triangular Cholesky decomposition of Σ_u.[2] It follows immediately from the condition $\Sigma_u = B_0^{-1} B_0^{-1\prime}$ that $B_0^{-1} = P$ is *one* possible solution to the problem of how to recover w_t. Since P is lower triangular, it has $K(K-1)/2$ zero parameters. As a result, the order condition for the exact identification of the unknown parameters in B_0^{-1} is satisfied. A useful result in this context is that B_0^{-1} being lower triangular implies that B_0 is lower triangular as well.

[1] The latter representation is sometimes referred to as the AB-model with $Au_t = Bw_t$, the model with $B = I$ as the A-model, and the model with $A = I$ as the B-model (see, e.g., Lütkepohl 2005, section 9.1). There is no universally accepted notation, however. Amisano and Giannini (1997), for example, refer to the AB-model, K-model and C-model.

[2] Standard software provides built-in functions for generating the Cholesky decomposition of Σ_u.

It is important to keep in mind that the orthogonalization of the reduced-form residuals by applying a Cholesky decomposition is appropriate only if the recursive structure embodied in P can be justified on economic grounds. The distinguishing feature of orthogonalization by Cholesky decomposition is that the resulting structural model is recursive (conditional on lagged variables). This means that we impose a particular causal chain rather than learning about the nature of the causal relationships from the data. In essence, we solve the problem of which structural shock causes the variation in u_t by imposing a particular solution. This mechanical solution does not make economic sense, however, without a plausible economic interpretation for the recursive ordering.

The neutral and scientific-sounding term "orthogonalization" hides the fact that we are making strong identifying assumptions about the error term of the VAR model. In the early 1980s, many users of VAR models did not appreciate this point and thought the data alone would speak for themselves. Such "atheoretical" VAR models soon were severely criticized (see, e.g., Cooley and Leroy 1985). This critique prompted more careful attention to the economic underpinnings of recursive models. It was shown that in special cases the recursive model can be given a structural or semistructural interpretation (e.g., Bernanke and Blinder 1992). This critique also spurred the development of structural VAR models that impose nonrecursive identifying restrictions (e.g., Sims 1986; Bernanke 1986; Blanchard and Watson 1986). It became widely recognized that the structural VAR model is simply a special case of the DSEM, the main distinguishing feature of the structural VAR model being the nature of its identifying restrictions.

In practice, there is a different solution P for each ordering of the K variables in the VAR model. It is sometimes argued that one should conduct sensitivity analysis based on alternative orderings of the K variables. This proposal is problematic for three reasons:

1. On the one hand, we claim to be sure that the ordering is recursive, yet on the other hand we claim to have no clue in what order the variables are recursive. This approach is not credible.
2. For a small VAR model with $K = 4$, for example, there are $4 \cdot 3 \cdot 2 \cdot 1 = 24$ permutations of the ordering. Few researchers would be willing to try out this many model specifications, nor would there be much hope that the results would be the same in each case unless the reduced-form errors are uncorrelated, which can be checked by inspecting the off-diagonal elements of Σ_u.
3. Even if there were no difference across these 24 specifications, this would only prove that the results are robust across all recursive orderings, but there is no reason for the true model to be recursive in the first place. This point is best illustrated by example. Let p_t denote the

price and q_t the quantity of a good. Price and quantity are driven by structural demand shocks w_t^d and supply shocks w_t^s. All dynamics are suppressed for expository purposes such that $y_t = u_t$:

$$\underbrace{\begin{pmatrix} p_t \\ q_t \end{pmatrix}}_{u_t} = \underbrace{\begin{bmatrix} 1 & -0.5 \\ 0.5 & 1 \end{bmatrix}}_{B_0^{-1}} \underbrace{\begin{pmatrix} w_t^d \\ w_t^s \end{pmatrix}}_{w_t}.$$

In this example, by construction, Σ_u is diagonal and the observable data are uncorrelated such that all recursive orderings are identical. This outcome obviously does not imply that any of the recursive orderings are valid. In fact, B_0^{-1} differs from

$$P = \text{chol}(\Sigma_u) = \text{chol}\left(\begin{bmatrix} 1.25 & 0 \\ 0 & 1.25 \end{bmatrix}\right) = \begin{bmatrix} 1.118 & 0 \\ 0 & 1.118 \end{bmatrix}$$

by construction. This point holds more generally. Let $u_t = B_0^{-1} w_t$ denote the true structural relationship and $u_t = P w_t^{chol}$ be the Cholesky relationship. Then

$$w_t^{chol} = P^{-1} u_t = P^{-1} \left(B_0^{-1} w_t\right) \neq w_t,$$

so the Cholesky decomposition will fail to identify the true structural shocks unless $B_0^{-1} = P$.

8.3 Sources of Identifying Restrictions

In the preceding subsection we stressed that unless we can come up with a convincing rationale for a particular recursive ordering, the resulting VAR impulse responses, forecast error variance decompositions, and historical decompositions are economically meaningless. This raises the question of where the economic rationale of the identifying restrictions on B_0^{-1} or B_0 comes from. There are a number of potential sources.

One is economic theory. In some cases, we may wish to impose the structure provided by a specific economic model. Of course, in that case the empirical results will only be as credible as the underlying model. A case in point is Blanchard's (1989) structural VAR analysis of the traditional Keynesian model involving an aggregate demand equation, Okun's law, a price-setting equation, the Phillips curve, and a monetary policy rule. Another example is Sims and Zha (2006a) who propose a VAR model with identifying assumptions that are approximately satisfied in a fully specified DSGE model. A third example is Fisher (2006) who derives his VAR identifying assumptions from a specific real business cycle model.

Another strategy is to specify an encompassing model that includes as special cases various alternative structural models implied by different economic models, allowing tests for overidentifying restrictions. The advantage of this approach is that it avoids conditioning on one specific model that may be incorrect. Of course, this type of structural VAR model no longer admits a Cholesky representation and must be estimated by alternative numerical methods (see Chapter 9). This strategy has been used, for example, by Bernanke and Mihov (1998b) who model the market for bank reserves as part of a study of U.S. monetary policy.

Often there is no fully developed theoretical model available, in which case identification may be achieved by using extraneous information or by using selective insights from economic theory:

Information delays. Information may not be available instantaneously because data are released only infrequently, allowing us to rule out instantaneous feedback. This approach has been employed by Inoue, Kilian, and Kiraz (2009), for example, who exploit differences in the timing of data releases to motivate the imposition of short-run exclusion restrictions. They observe that Survey of Professional Forecasters (SPF) inflation forecasts for the current quarter are formed before the interest rate for the current quarter is set (reflecting their release in the middle of the preceding quarter), while household survey expectations of the real interest rate are formed after observing the nominal interest rate at the beginning of the quarter. This reasoning suggests a recursive structure with the SPF forecast ordered first, followed by the interest rate and households' real interest rate expectations. This structure allows us to assess the impact of inflation news contained in the SPF forecasts on household expectations of the inflation rate. A similar argument is invoked by Sims (1998) who suggests that monetary policymakers react immediately only to variables that they can observe without delay (such as commodity prices, monetary aggregates, and financial variables), but not to variables that they can observe only with a delay (such as real GDP or the GDP deflator).

Physical constraints. For example, it takes time for a firm's investment decision to be made and for the new equipment to be installed, so measured physical investment responds only with a delay to investment decisions. This delay motivates an exclusion restriction of the impact response of investment to new information about the economy. This type of friction is consistent with the use of time-to-build technologies in many DSGE models. More generally, it is common to postulate that variables such as real GDP, industrial production, and real investment are inherently sluggish and hence do not respond contemporaneously to monetary policy shocks (see Sims 1998).

Institutional knowledge. For example, we may have information about the inability of suppliers to respond to demand shocks in the short run due to adjustment costs, which amounts to imposing a vertical slope on the short-run supply curve (see, e.g., Kilian 2009). Similarly, Davis and Kilian (2011) exploit the fact that gasoline taxes (excluding ad valorem taxes) do not respond instantaneously to the state of the economy because lawmakers move at a slow pace.

Assumptions about market structure. Another common identifying assumption in empirical work is that there is no feedback from a small open economy to the rest of the world. This identifying assumption has been used, for example, to motivate treating U.S. interest rates as contemporaneously exogenous with respect to the macroeconomic aggregates of small open economies such as Canada (see, e.g., Cushman and Zha 1997).

Homogeneity of demand functions. Another possible source of identifying information are homogeneity restrictions on demand functions. For example, Galí (1992) imposes short-run homogeneity on the demand for money when assuming that the demand for real balances is not affected by contemporaneous changes in prices (given the nominal rate and output). This assumption amounts to assuming away costs of adjusting nominal money holdings. Similar homogeneity restrictions have also been used in Bernanke (1986). Likewise, Keating (1992) postulates that money holdings rise in proportion to nominal income such that the response to a change in real income is the same as the response to a change in the price level.

Extraneous parameter estimates. When the elements of B_0 can be viewed as elasticities, it may be possible to impose values for those elasticities based on extraneous information from other studies. This approach has been used by Blanchard and Perotti (2002), for example. Similarly, Blanchard and Watson (1986) impose nonzero values for some structural parameters in B_0 based on extraneous information. If the parameter value cannot be pinned down with any degree of reliability, yet another possibility is to explore a grid of possible structural parameter values, as in Abraham and Haltiwanger (1995) and Blanchard (1989). A similar approach has also been used in Kilian (2010) and Davis and Kilian (2011) in an effort to assess the robustness of their baseline results. In a different context, Todd (1990) interprets Sims' (1980b) recursive VAR model of monetary policy in terms of alternative assumptions about the slopes of money demand and money supply curves.

High-frequency data. In rare cases, it may be possible to motivate exclusion restrictions more directly. For example, Kilian and Vega (2011) use daily data on U.S. macroeconomic news to formally test the identifying assumption of no feedback within the month from U.S. macroeconomic aggregates to the dollar price of oil and the U.S. price of gasoline. They demonstrate that this test is unable to reject the null of no feedback for these energy prices, while having enough power to reject the same null for a large number of asset prices. This result lends credence to exclusion restrictions in monthly VAR models ruling out instantaneous feedback from U.S. macroeconomic aggregates to the U.S. prices of oil and gasoline.

It is fair to say that coming up with a set of credible short-run identifying restrictions is difficult. Whether a particular exclusion restriction is convincing often depends on the data frequency, and in many cases there are not enough credible exclusion restrictions to achieve full identification. This fact has stimulated interest in the alternative identification methods discussed in subsequent chapters.

8.4 Examples of Recursively Identified Models

8.4.1 A Simple Macroeconomic Model

Let $y_t = (p_t, gdp_t, m_t, i_t)'$, where p_t is the log price level, gdp_t is log real GDP, m_t the log of a monetary aggregate such as M1, and i_t the federal funds rate. The data are quarterly and the proposed identification is recursive such that

$$\begin{pmatrix} u_t^p \\ u_t^{gdp} \\ u_t^m \\ u_t^i \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} & 0 \\ b_0^{41} & b_0^{42} & b_0^{43} & b_0^{44} \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{pmatrix}. \tag{8.4.1}$$

Note that each line can be viewed as an equation. Each reduced-form shock is a weighted average of selected structural shocks. The letters b_0^{ij} represent the weights attached to the structural shocks. For example, the first equation is $u_t^p = b_0^{11} w_{1t} + 0 + 0 + 0$, the second reads $u_t^{gdp} = b_0^{21} w_{1t} + b_0^{22} w_{2t} + 0 + 0$, etc.

One way of rationalizing this identification would be to interpret the first two equations as an aggregate supply (AS) and aggregate demand (AD) model with a horizontal AS curve and downward-sloping AD curve. w_{1t} moves the price level and real output, so it must be a shift of the AS curve. w_{2t} moves real output only, so it must represent a shift of the AD curve. The third equation could be interpreted as a money demand equation derived from the quantity equation: $MV = PY$, where V stands for velocity and Y for real income. Hence, w_{3t} can be interpreted as a velocity shock or money demand shock, if we take real GDP to represent real income. The last equation could represent a monetary policy reaction function. The Federal Reserve systematically responds to u_t^p, u_t^{gdp}, and u_t^m (as well as lags of all variables). Any change in the interest rate not accounted for by this response would be an exogenous monetary policy (or money supply) shock. Such policy shocks could arise from changes in the composition of the Federal Open Market Committee, for example, or may reflect reactions to idiosyncratic events such as the 9/11 terrorist attack on the United States or the housing crisis that are not captured by standard policy rules.

It is easy to spot the limitations of this model. For example, why does money demand not respond to the interest rate within a quarter? How plausible is the assumption of a horizontal short-run supply curve? These are the types of questions that one must ask when assessing the plausibility of a structural VAR model. This example also illustrates that theory typically is not sufficient for identification, even if we are willing to condition on a particular theoretical model. For example, if the AS curve were vertical, but the AD curve horizontal by assumption, the first two equations of the structural model (8.4.1) would have to be modified. More generally, no recursive structure would be able to accommodate a theoretical model in which the AS and AD curves are neither horizontal nor vertical, but upward and downward sloping. This point highlights the difficulty of specifying fully structural models of the macroeconomy in recursive form and explains why such models have declined in importance in macroeconomic analysis.

8.4.2 A Model of the Global Market for Crude Oil

The second example is a structural VAR model of the global market for crude oil based on Kilian (2009). Let $y_t = (\Delta prod_t, rea_t, rpoil_t)'$, where $\Delta prod_t$ denotes the percent change in world crude oil production, rea_t is a business cycle index measuring global real economic activity, and $rpoil_t$ is the log of the real price of oil. The data are monthly. Let

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix} \begin{pmatrix} w_t^{\text{oil supply}} \\ w_t^{\text{aggregate demand}} \\ w_t^{\text{oil-specific demand}} \end{pmatrix}.$$

This model of the global market for crude oil embodies a vertical short-run oil supply curve and a downward-sloping short-run oil demand curve (conditional on lags of all variables). There are two demand shocks that are separately identified by the delay restriction that oil-specific demand shocks raise the real price of oil, but without affecting global real economic activity within the same month, while shocks to the aggregate demand for all industrial commodities affect both real activity and the real price of oil on impact.

One might question whether one could have imposed an overidentifying restriction of the form $b_0^{21} = 0$. In other words, one would expect that higher oil prices triggered by unanticipated oil supply disruptions would not slow down global real activity within the month any more or less than oil-specific demand shocks. It turns out that the estimate of b_0^{21} is essentially zero, even without imposing that restriction, making this point moot.[3] One could also question whether the short-run supply curve is truly vertical. Defending this assumption requires institutional knowledge of oil markets or extraneous econometric evidence. Additional theoretical support for this assumption is provided in Anderson, Kellogg, and Salant (2016).

Kilian (2009) constructs historical decompositions based on this VAR model that establish that oil demand shocks rather than oil supply shocks explain most fluctuations in the real price of oil, not just in recent years, but as far back as in the late 1970s. Moreover, structural impulse response analysis reveals that no two oil price surges are alike. Depending on the composition of the underlying oil demand and oil supply shocks, the persistence, shape, and magnitude of the response of the real price of oil may differ substantially. This result has important implications for studies of the transmission of oil price shocks.

8.4.3 Oil Price Shocks and Stock Returns

Kilian and Park (2009) use a generalization of the model in Kilian (2009) to study the dynamic response of U.S. real stock prices to oil supply shocks, aggregate demand shocks, and oil-specific demand shocks in the global market for crude oil:

$$\begin{pmatrix} u_t^{\Delta prod} \\ u_t^{rea} \\ u_t^{rpoil} \\ u_t^{ret} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} & 0 \\ b_0^{41} & b_0^{42} & b_0^{43} & b_0^{44} \end{bmatrix} \begin{pmatrix} w_t^{\text{oil supply}} \\ w_t^{\text{aggregate demand}} \\ w_t^{\text{oil-specific demand}} \\ w_t^{\text{other}} \end{pmatrix},$$

where ret_t denotes real U.S. stock returns and u_t^{ret} is the reduced-form error of the corresponding VAR equation. In this model, U.S. real stock returns may respond instantaneously to any of the shocks that determine the real price of oil, as one would expect of an asset price. The residual shock w_t^{other} collectively captures all shocks that affect U.S. real stock returns without affecting the real price of oil within the same month.

The key additional identifying assumption is that the real global price of oil does not respond within the current month to changes in U.S. real stock returns that are not already explained by the first three shocks. This assumption allows for expectations of a booming world economy to drive up both the global real price of oil and U.S. real stock returns. It does not allow for an unrelated U.S. stock market correction to affect the real price of oil within the same month. This identifying assumption is implied by the conventional assumption that the

[3] Further discussion of this point can be found in Chapters 9 and 12.

real price of oil is predetermined with respect to U.S. macroeconomic aggregates because

$$u_t^{rpoil} = b_0^{31} w_t^{\text{oil supply}} + b_0^{32} w_t^{\text{aggregate demand}} + b_0^{33} w_t^{\text{oil-specific demand}}.$$

By cumulating the return responses, one can infer the response of U.S. real stock prices to each oil demand and oil supply shock. Kilian and Park illustrate that the dynamic correlation between oil returns and stock returns evolves with changes in the composition of oil demand and oil supply shocks and may be positive, negative, or zero. This finding explains the seeming instability in the reduced-form relationship between oil prices and stock prices documented in the finance literature.

8.4.4 Models of the Transmission of Energy Price Shocks

Recursively identified VAR models such as the model in Kilian (2009) are fully identified in that each structural shock is uniquely identified. Often we do not have enough restrictions to fully identify a VAR model. This has prompted the development of semistructural or partially identified VAR models. The idea of semistructural models is that in some cases we may be satisfied if we can identify a subset of the structural shocks. Often we are interested in one structural shock only. An example is models of the transmission of energy price shocks in which the price of energy is predetermined with respect to all domestic macroeconomic aggregates, consistent with the empirical evidence provided in Kilian and Vega (2011). For example, Edelstein and Kilian (2009) utilized a recursively identified monthly bivariate model similar to the model:

$$\begin{pmatrix} u_t^{\Delta p} \\ u_t^{\Delta c} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 \\ b_0^{21} & b_0^{22} \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix},$$

where Δp_t denotes the percent change in U.S. energy prices (possibly weighted by the energy share in expenditures) and Δc_t denotes the percent growth in real U.S. consumption. The model is semistructural in that only the innovation in the price of energy, w_{1t}, is economically identified. The w_{2t} term, in contrast, is a conglomerate of other structural shocks that are not individually identified. Hence, w_{2t} does not have an economic interpretation. Put differently, both elements of w_t are statistically identified, but only w_{1t} is economically identified. The identification scheme postulates that Δp_t is predetermined with respect to Δc_t.

It is important to keep in mind that responses to energy price innovations from models such as this one are designed to tell us about the average response to an energy price shock over the sample. The actual response at a given point in time may be different from the average response, depending on the determinants of the energy price shock in question. Nevertheless, models of this type have been used extensively in the literature, given the challenges of identifying energy demand and supply shocks (see Rotemberg and Woodford 1996).

8.4.5 Semistructural Models of Monetary Policy

A large literature is devoted to identifying monetary policy shocks using structural VAR models that are only partially identified. Under suitable conditions such models may be estimated using standard techniques for recursively identified VAR models with the monetary policy instrument ordered last among the VAR variables. This idea was pioneered by Bernanke and Blinder (1992).

A Stylized Model of U.S. Monetary Policy. The simplest example is a quarterly model for $y_t = (\Delta gdp_t, \pi_t, i_t)'$, where Δgdp_t denotes U.S. real GDP growth, π_t the inflation rate, and i_t the federal funds rate. We use the lower-triangular Cholesky decomposition to compute

$$\begin{pmatrix} u_t^{\Delta gdp} \\ u_t^{\pi} \\ u_t^{i} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 & 0 \\ b_0^{21} & b_0^{22} & 0 \\ b_0^{31} & b_0^{32} & b_0^{33} \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \end{pmatrix}.$$

The last equation of the model is interpreted as a linear monetary policy reaction function. The interest rate is the policy instrument. In setting u_t^i, the Federal Reserve responds endogenously to contemporaneous variation in Δgdp_t and π_t and to variation in the lagged observables. This part of the model captures the systematic response of the central bank to the economy. The residual left after accounting for all endogenous variation in the interest rate, w_{3t}, is interpreted as an exogenous monetary policy shock. This policy shock may reflect deviations from the expected (or average) policy response that may arise, for example, from changes in the composition of the Federal Open Market Committee (FOMC), from shifts in the weights that the FOMC attaches to different objectives, or from discretionary policy decisions in response to extraordinary events. It may also arise from self-fulfilling shocks to market expectations about Fed policy, or from errors the Federal Reserve makes in forecasting the extent to which Treasury operations will add or drain reserves available to private banks (see, e.g., Chari, Christiano, and Eichenbaum 1998; Hamilton 1997). The policy shock, w_{3t}, is the only structural shock of interest in this model. No attempt is made to identify economically the structural shocks w_{1t} and w_{2t}.

It is easy to see that alternative orderings of $u_t^{\Delta gdp}$ and u_t^{π} will leave w_{3t} unaffected because the last column of B_0^{-1} is unaffected by the elements in the first two columns of this matrix. This result even allows for a nonrecusive relationship between $u_t^{\Delta gdp}$ and u_t^{π}. A related result about the invariance of the monetary policy shock to alternative block-triangular orderings of B_0 can be found in Christiano, Eichenbaum, and Evans (1999).

Models of this type have been commonly used in empirical work. The policy variable in semistructural VAR models need not be the short-term interest rate. A similar approach to identification may be followed with alternative policy indicators such as nonborrowed reserves (see, e.g., Strongin 1995). Regardless of the details of the specification, this identification scheme requires that the shock of interest be ordered at (or near) the bottom of the recursive ordering, given that the central bank tends to respond to most (if not all) of the other model variables.

Semistructural VAR models of monetary policy have five important weaknesses that are apparent even in the trivariate setting. First, this model does not allow for feedback within a given quarter from w_{3t} to Δgdp_t and π_t. This restriction seems implausible at least at quarterly frequency. Because Δgdp_t is not available at higher frequency, there is little we can do about this problem.[4] It might seem that the same identification scheme would be more credible if we replaced Δgdp_t by the growth rate of industrial production and estimated the model at monthly frequency. This is not the case. One problem is that industrial output accounts for only a fraction of total output and that fraction is unstable over time. Moreover, real GDP is a measure of value added, whereas industrial output is a gross output measure. Finally, it is well known that the Federal Reserve is concerned with broader measures of real activity, making a policy reaction function based on industrial production growth economically less plausible and hence less interesting. In this regard, a better measure of monthly U.S. real activity would be the Chicago Fed's monthly principal components index of U.S. real activity (CFNAI). Yet another approach in the literature has been to interpolate quarterly real GDP data based on the fluctuations in monthly industrial production data and other monthly indicators. Such ad hoc methods not only suffer from the same deficiencies as the use of industrial production data, but they are also likely to distort the structural impulse responses to be estimated.

Second, the Federal Reserve may respond systematically to more variables than just Δgdp_t and π_t. To the extent that these variables are omitted from the model, we obtain inconsistent estimates of b_0^{31} and b_0^{32}, and incorrect measures of the monetary policy shock w_{3t}. In essence, the problem is that the policy shocks must be exogenous to allow us to learn about the effects of monetary policy shocks.

One potential solution to this problem is to enrich the set of variables ordered above the interest rate relative to the simple benchmark model and to estimate much larger VAR systems (see, e.g., Bernanke and Blinder 1992;

[4] The Bureau of Economic Analysis does not release monthly U.S. real GDP data. Unofficial measures of monthly U.S. real GDP constructed similarly to the official quarterly data have recently been provided by Macroeconomic Advisers, LLC. These time series currently are not long enough for estimating VAR models of monetary policy, however.

Sims 1992; Christiano, Eichenbaum, and Evans 1999). Adding more variables, however, invites overfitting and undermines the credibility of the VAR estimates. Standard VAR models cannot handle more than about half a dozen variables, given typical sample sizes.

One potential remedy of this problem is to work with factor-augmented VAR (FAVAR) models, as in Bernanke and Boivin (2003), Bernanke, Boivin, and Eliasz (2005), Stock and Watson (2005), or Forni, Giannone, Lippi, and Reichlin (2009). Alternatively, one can work with large-scale Bayesian VAR models in which the cross-sectional dimension K is allowed to be larger than the time dimension T, as in Bańbura, Giannone, and Reichlin (2010). These large-scale models are designed to incorporate a much richer information structure than conventional semistructural VAR models of monetary policy. FAVAR models and large-scale BVAR models have three distinct advantages over conventional small to medium-sized VAR models: (1) They allow for the fact that central bankers form expectations about domestic real activity and inflation based on hundreds of economic and financial time series rather than a handful of time series; (2) they allow for the fact that economic concepts such as domestic economic activity and inflation may not be well represented by a single observable time series; (3) they allow the user to construct the responses of many variables not included in conventional VAR models. There is evidence that allowing for richer information sets in specifying VAR models may improve the plausibility of the estimated responses. A detailed review of structural FAVAR models and large-scale structural BVAR models can be found in Chapter 16.

An alternative strategy is to add to the model selectively one or more variables that carry information about future economic activity and inflation and hence help address the informational deficiencies of low-dimensional VAR models (see Sims and Zha 2006a). We already noted in Chapter 7 that asset prices in particular quickly respond to new information about future economic conditions. Examples are housing prices, stock prices, and industrial commodity prices. Of particular interest in this context is the inclusion of an index of commodity prices.

A common problem in monetary policy VAR models has been the presence of a so-called price puzzle, which refers to the finding of an increase in the price level in response to an unanticipated monetary tightening. This finding suggests that the VAR model is unable to capture expected inflation. Sims (1992) suggested that this puzzle can be resolved by including global commodity prices as an indicator of expected inflation in the model.[5]

[5] Sims' premise was that the Federal Reserve considers global commodity prices as a predictor of expected inflation. Subsequently, Hanson (2004) showed that there is no systematic relationship between the ability of alternative measures of global commodity prices to predict U.S. inflation and their ability to resolve the price puzzle, casting doubt on this premise.

Following Sims, it has been customary to order commodity prices above the interest rate in the monetary policy VAR model, which allows the central bank to respond to movements in commodity prices. This assumption, however, prevents commodity prices from responding within the current period to policy changes in the interest rate. The latter restriction does not seem economically plausible. Ordering the commodity price index last addresses the latter concern, but is inconsistent with Sims' premise that the central bank responds contemporaneously to commodity price fluctuations. Thus, neither specification is economically appealing. Indeed, subsequent research has shown that the price puzzle more often than not persists even after including global commodity prices in the VAR model, suggesting that the model remains misspecified.

Third, the identification of the VAR model hinges on the monetary policy reaction function being stable over time. To the extent that policymakers have at times drastically changed the weights attached to their inflation and output objectives or have even changed the policy instrument, it becomes essential to split the sample in estimating the VAR model. The resulting shorter sample in turn makes it more difficult to include many variables in the model due to the lack of degrees of freedom. It also complicates statistical inference.

Fourth, the VAR model is linear. It does not allow for a lower bound on the interest rate, for example, making this model unsuitable for studying the quantitative easing of the Federal Reserve Board in recent years. Lower bounds on the interest rate imply a nonlinearity in the model. Nonlinear VAR models are discussed in Chapter 18.

Fifth, most VAR models of monetary policy ignore the real-time nature of the policy decision problem. Not all data relevant to policymakers are available without delay, and when data become available, they tend to be preliminary and subject to further revisions. To the extent that monetary policy shocks are defined as the residual of the policy reaction function, a misspecification of the policymaker's information set will cause biases in the estimated policy shocks. Bernanke and Boivin (2003) is an example of a study that explores the role of real-time data limitations in semistructural VAR models. Their conclusion is that – at least for their sample period and model – the distinction between real-time data and ex-post revised data is of limited importance. A similar conclusion is reached by Croushore and Evans (2006).

Finally, it is useful to reiterate that the thought experiment contemplated in structural VAR models is an unanticipated monetary policy shock within an existing monetary policy rule (also referred to as a policy regime at times). This exercise is distinct from that of changing the monetary policy rule (as happened in 1979 under Paul Volcker or in 2008 following the quantitative easing of the Federal Reserve Board). The question of what the effects of a shift in the monetary policy rule are on the economy is of independent interest, but much harder to answer. The role of systematic monetary policy has been stressed in Leeper, Sims, and Zha (1996), Bernanke, Gertler, and Watson

(1997), Hamilton and Herrera (2004), and Kilian and Lewis (2011), for example. The econometric evaluation of the role of systematic monetary policy in linear VAR models, however, remains controversial and may easily run afoul of the Lucas critique, as discussed in Chapter 4. One potential alternative is the use of nonlinear structural VAR models that allow for time variation, as discussed in Chapter 18.

Three Benchmark Models of U.S. Monetary Policy. Recursively identified models of monetary policy in practice tend to include many more variables than the stylized three-variable model we considered so far. For example, Christiano, Eichenbaum, and Evans (1999) propose three benchmark specifications for semistructural VAR models of monetary policy, which differ depending on whether we view the federal funds rate (i_t), nonborrowed reserves (nbr_t) or the ratio of nonborrowed to total reserves (nbr_t/tr_t) as the policy instrument. The models are quarterly and involve seven variables each. The first specification is

$$y_t = \begin{pmatrix} y_{1t} \\ i_t \\ y_{2t} \end{pmatrix},$$

where y_{1t} is a vector of macroeconomic aggregates that the Federal Reserve responds to contemporaneously in setting the policy instrument, and y_{2t} is a vector of macroeconomic aggregates that the Federal Reserve does not respond to within the current quarter. Whereas y_{1t} is assumed not to respond to policy surprises within the first quarter, y_{2t} is allowed to respond within the first quarter. Christiano et al. define $y_{1t} = (gdp_t, p_t, pcom_t)'$, where gdp_t denotes U.S. real GDP, p_t the GDP deflator, and $pcom_t$ a commodity price index, and $y_{2t} = (tr_t, nbr_t, m_t)'$, where tr_t and nbr_t denote total reserves and nonborrowed reserves, respectively, and m_t refers to a suitably defined monetary aggregate. All variables but the interest rate are expressed in logs.

The second specification is

$$y_t = \begin{pmatrix} y_{1t} \\ nbr_t \\ y_{2t} \end{pmatrix},$$

where the policy instrument nbr_t is nonborrowed reserves, $y_{1t} = (gdp_t, p_t, pcom_t)'$, and $y_{2t} = (i_t, tr_t, m_t)'$. The third specification is

$$y_t = \begin{pmatrix} y_{1t} \\ nbr_t \\ y_{2t} \end{pmatrix},$$

where again the policy instrument nbr_t is nonborrowed reserves, $y_{1t} = (gdp_t, p_t, pcom_t, tr_t)'$, and $y_{2t} = (i_t, m_t)'$. Christiano et al. refer to the resulting policy shock as an nbr_t/tr_t shock, building on Strongin (1995) who proposed the ratio of nonborrowed reserves to total reserves as an alternative monetary policy instrument.

There is a large literature on the theoretical merits of each of these specifications and on how to model the relationship between the federal funds rate, nonborrowed reserves, and total reserves. A review of this literature is provided in Christiano, Eichenbaum, and Evans (1999), who also demonstrate that, in all three specifications, the federal funds rate increases, monetary aggregates decline over time, the price level initially is sluggish, real output temporarily declines, and commodity prices fall in response to an unanticipated monetary contraction.

A Model of the Macroeconomic Effects of Monetary Policy. More recent models in this literature have tended to ignore data on nonborrowed reserves and total reserves and have largely focused on the federal funds rate as the policy instrument. For example, Christiano, Eichenbaum, and Evans (2005) analyze a nine-variable version of the recursive VAR model of monetary policy. The data are quarterly. As before,

$$y_t = \begin{pmatrix} y_{1t} \\ i_t \\ y_{2t} \end{pmatrix},$$

where i_t is the federal funds rate. Christiano et al. define

$$y_{1t} = (gdp_t, c_t, p_t, inv_t, wage_t, prod_t)',$$

where gdp_t denotes U.S. real GDP, p_t the GDP deflator, c_t real consumption, inv_t real investment, $wage_t$ the real wage, and $prod_t$ labor productivity. They define $y_{2t} = (\Delta m_t, rp_t)'$, where Δm_t refers to M2 growth and rp_t stands for real profits. All variables but the interest rate and Δm_t are expressed in logs. The monetary policy shock is identified by imposing a recursive structure on B_0^{-1} (and hence on B_0) such that the seventh row of B_0 may be interpreted as a policy reaction function conditional on past data of all model variables. None of the other structural shocks are identified from an economic point of view.

The Effect of Monetary Policy Shocks on the Exchange Rate. Eichenbaum and Evans (2005) extend the closed-economy semistructural VAR framework to include the bilateral U.S. dollar exchange rate with respect to selected other countries. The data are monthly. Let

$$y_t = \begin{pmatrix} y_{1t} \\ i_t^{\mathrm{US}} \\ y_{2t} \end{pmatrix},$$

where i_t^{US} is the U.S. federal funds rate. Eichenbaum and Evans define

$$y_{1t} = \left(ip_t^{\mathrm{US}}, p_t^{\mathrm{US}}\right)',$$

where ip_t^{US} is the log of U.S. industrial production and p_t^{US} denotes the log of the U.S. consumer price index, and

$$y_{2t} = \begin{pmatrix} i_t^{\mathrm{Foreign}} - i_t^{\mathrm{US}} \\ er_t^{\mathrm{US/Foreign}} \end{pmatrix},$$

where $i_t^{\mathrm{Foreign}} - i_t^{\mathrm{US}}$ is the short-term nominal interest rate differential and $er_t^{\mathrm{US/Foreign}}$ denotes, alternatively, the U.S. nominal exchange rate or the U.S. real exchange rate. The foreign countries (considered one at a time) include Japan, Germany, Italy, France, and the United Kingdom. One identifying assumption is that the Federal Reserve responds to the current level of U.S. real output and U.S. prices but disregards the current foreign interest rate and the current exchange rate in setting policy. The other identifying assumption is that U.S. real output and the U.S. price level do not respond within the month to monetary policy disturbances. The model is again partially identified in that only the monetary policy shock is economically identified.

The study focuses on two questions. One is the exchange rate responses to a monetary policy shock. Eichenbaum and Evans document a persistent appreciation of nominal and real exchange rates in response to a contractionary monetary policy shock. The other question is whether uncovered interest rate parity applies to the responses to monetary policy shocks. Based on the estimated model, Eichenbaum and Evans compute the implied excess returns defined as the ex post difference in the return between investing \$1 in one-period foreign bonds and investing \$1 in one-period domestic bonds, expressed in U.S. dollars. Under uncovered interest rate parity, one would not expect a monetary policy shock to be associated with excess returns. Eichenbaum and Evans find persistent and statistically significant departures from uncovered interest rate parity.

8.4.6 The Permanent Income Model of Consumption

Cochrane (1994) proposes another application of the recursive model. His interest is not in identifying demand or supply shocks, but in decomposing permanent and transitory shocks within the framework of the permanent income model of consumption. The standard permanent income model implies that log real consumption (c_t) and log real income (gnp_t) are cointegrated such that the consumption-income ratio is stationary. Cochrane imposes this cointegration restriction on the reduced-form VAR model for (c_t, gnp_t). The permanent income model also predicts that if income changes unexpectedly without a corresponding change in consumption, then consumers will regard the shock to income as having purely transitory effects on income. Cochrane identifies such a shock by recursively ordering innovations to consumption first in the Cholesky decomposition of the reduced-form error covariance matrix. This decomposition allows him to separate permanent from transitory shocks and to quantify their importance for the variability of consumption and income:

$$\begin{pmatrix} u_t^c \\ u_t^{gnp} \end{pmatrix} = \begin{bmatrix} b_0^{11} & 0 \\ b_0^{21} & b_0^{22} \end{bmatrix} \begin{pmatrix} w_t^{\text{permanent}} \\ w_t^{\text{transitory}} \end{pmatrix}.$$

Note that, by construction, consumption only depends on the permanent shock, whereas income in addition depends on the transitory shock.[6] Cochrane verifies that the response of income to the transitory shock is indeed rapidly mean-reverting, whereas the response of income to a shock that moves both consumption and income on impact has long-lasting effects on income, as expected from a permanent shock. Moreover, much of the consumption response to a permanent shock is immediate, whereas the response of consumption to a transitory shock is close to zero at all horizons.[7]

Unlike in our earlier examples, this methodology is silent about the economic interpretation of permanent and transitory shocks. It is common in the literature, however, to equate the permanent shock in this model with a supply or productivity shock and the transitory shock with a demand shock. In general, the transitory and permanent shocks will be a mixture of these deeper economic shocks.

8.5 Examples of Nonrecursively Identified Models

Not all structural VAR models have a recursive structure. Increasing skepticism toward atheoretical recursively identified models in the mid-1980s stimulated a series of studies proposing explicitly structural models identified by nonrecursive short-run restrictions (see, e.g., Bernanke 1986; Sims 1986; Blanchard and Watson 1986). As in the recursive model, the identifying restrictions on B_0 or B_0^{-1} generate moment conditions that can be used to estimate the unknown coefficients in B_0. In general, solving the moment conditions for the unknown structural parameters will require iteration, but in some cases the GMM estimator can be constructed using traditional instrumental-variable techniques. An alternative commonly used approach is to model the error distribution as Gaussian and to estimate the structural model by full information maximum

6 The terminology of transitory shocks and permanent shocks is somewhat misleading in that any shock by construction involves a one-time disturbance only. A transitory shock, more precisely, is defined as a shock with purely transitory effects on the observables, whereas a permanent shock refers to a shock with permanent (or long-run) effects on the observables.

7 It can be shown that the results of Cochrane's model would be identical to the results from a model in which the transitory shock has no long-run effect on the level of income and consumption. Such long-run restrictions are discussed in Chapter 10.

likelihood methods. This approach involves the maximization of the concentrated likelihood with respect to the structural model parameters subject to the identifying restrictions. A detailed discussion of these estimation issues is provided in Chapter 9. Our focus in this section is on providing examples of identifying assumptions used in this literature.

8.5.1 Fiscal Policy Shocks

Blanchard and Perotti (2002) introduce a model of U.S. fiscal policy that deviates from the usual recursive structure. They propose a quarterly model of the U.S. economy for $y_t = (tax_t, gov_t, gdp_t)$, where tax_t refers to real taxes, gov_t to real government spending, and gdp_t to real GDP. All variables are in logs. Ignoring lags, the structural relations can be written as

$$\begin{bmatrix} 1 & 0 & b_{13,0} \\ 0 & 1 & b_{23,0} \\ b_{31,0} & b_{32,0} & 1 \end{bmatrix} \begin{pmatrix} u_t^{tax} \\ u_t^{gov} \\ u_t^{gdp} \end{pmatrix} = \begin{bmatrix} 1 & c_{12} & 0 \\ c_{21} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} w_t^{tax} \\ w_t^{gov} \\ w_t^{gdp} \end{pmatrix}$$

such that

$$\begin{aligned} u_t^{tax} &= -b_{13,0} u_t^{gdp} + c_{12} w_t^{gov} + w_t^{tax}, \\ u_t^{gov} &= -b_{23,0} u_t^{gdp} + c_{21} w_t^{tax} + w_t^{gov}, \\ u_t^{gdp} &= -b_{31,0} u_t^{tax} - b_{32,0} u_t^{gov} + w_t^{gdp}. \end{aligned}$$

Blanchard and Perotti first provide institutional arguments for the delay restriction $b_{23,0} = 0$ which rules out automatic feedback from economic activity to government spending within the quarter. They then show that the within-quarter response of taxes to economic activity, $-b_{13,0}$, can be derived on the basis of extraneous tax elasticity estimates and is $b_{13,0} = -2.08$. The parameters $b_{31,0}$ and $b_{32,0}$ are left unrestricted. The potential endogeneity between taxes and spending is dealt with by imposing either $c_{21} = 0$ or $c_{12} = 0$. In the latter case, for example, we obtain

$$\begin{aligned} u_t^{tax} &= 2.08 u_t^{gdp} + w_t^{tax}, \\ u_t^{gov} &= c_{21} w_t^{tax} + w_t^{gov}, \\ u_t^{gdp} &= -b_{31,0} u_t^{tax} - b_{32,0} u_t^{gov} + w_t^{gdp}. \end{aligned}$$

This system can be estimated numerically, as discussed in Chapter 9. Note that Blanchard and Perotti effectively treat the first two innovations as mutually exogenous without imposing the overidentifying restriction on c_{21}. An obvious concern is that the model does not allow for the anticipation of fiscal shocks. Blanchard and Perotti discuss how this concern may be addressed by changing the timing assumptions and adding further identifying restrictions, if we are willing to postulate a specific form of foresight. Another concern is that the model does not condition on the debt structure (see, e.g., Chung and Leeper 2007). Allowing for the debt structure to matter would result in a nonlinear dynamic model.

8.5.2 An Alternative Simple Macroeconomic Model

Keating (1992) discusses a variation of the simple macroeconomic model we discussed earlier that does not impose a recursive structure and involves a different economic interpretation. His structural equations are

$$\begin{aligned} u_t^p &= w_t^{AS}, \\ u_t^{gdp} &= -b_{21,0}u_t^p - b_{23,0}u_t^i - b_{24,0}u_t^m + w_t^{IS}, \\ u_t^i &= -b_{34,0}u_t^m + w_t^{MS}, \\ u_t^m &= -b_{42,0}(u_t^{gdp} + u_t^p) - b_{43,0}u_t^i + w_t^{MD}. \end{aligned}$$

The first equation again represents a horizontal AS curve, but the second equation now can be interpreted as an IS curve, allowing real output to respond to all other model variables. The third equation represents a simple money supply function, according to which the central bank adjusts the rate of interest in relation to the money stock, and the fourth equation is a money demand function in which short-run money holdings rise in proportion to nominal income, yielding the final restriction required for exact identification. Unlike in the earlier example, money holdings are allowed to depend on the interest rate as well. Clearly, this model specification embodies a very different view of what monetary policymakers do than more recently developed structural VAR models motivated by the literature on Taylor rules (see Taylor 1993).

8.5.3 Discussion

Nonrecursively identified VAR models more closely resemble traditional simultaneous equations models. This means that they also are susceptible to the weaknesses of such models, including the difficulty of finding strong instruments for identifying causal effects in the data (see Chapter 9). A case in point is the literature on the liquidity effect. The liquidity effect refers to the short-run negative response of interest rates to an unanticipated monetary expansion. Although the presence of such an effect has been suspected for a long time, it has only been in the 1990s that structural VAR studies emerged concluding that there is a liquidity effect. Whereas the evidence of a liquidity effect is at best mixed in recursively identified models of monetary policy, empirical VAR studies based on nonrecursive simultaneous equations systems have reliably produced a strong liquidity effect. This evidence might seem to suggest that more explicitly structural models are inherently superior to earlier semistructural models of monetary policy. However, Pagan and Robertson (1998) show that the instruments underlying the three most important nonrecursive studies of the liquidity effect appear weak in the econometric sense, calling into question any inferences made about the magnitude of the liquidity effect. We defer the discussion of alternative estimation methods for recursively and nonrecursively identified structural VAR models to Chapter 9.

8.5.4 The Graph-Theoretic Approach

Another strand of the literature uses graph-theoretic tools and terminology for representing causal relations between variables (see Pearl 2000; Spirtes, Glymour, and Scheines 2000; Eichler 2012). The idea is to represent causal relations between the variables of interest by directed graphs and to use statistical methods to narrow down the admissible causal structures that are compatible with the data. As a result, this approach has become known as the graph-theoretic approach to identification. The graph-theoretic approach was first used for structural VAR analysis by Swanson and Granger (1997) and was later applied, extended, and refined by Bessler and Lee (2002), Demiralp and Hoover (2003), Hoover (2005), Demiralp, Hoover, and Perez (2008), Hoover, Demiralp, and Perez (2009), Moneta (2008), Moneta, Entner, Hoyer, and Coad (2013), and Heinlein and Krolzig (2013), among others.

The graph-theoretic approach explores the partial correlations between the reduced-form VAR model errors. For example, in a three-dimensional Gaussian model, if the partial correlation between u_{1t} and u_{3t}, given u_{2t}, is zero, u_{1t} and u_{3t} are independent conditionally on u_{2t}. Hence, the coefficient of u_{3t} in a regression

$$u_{1t} = -b_{12,0}u_{2t} - b_{13,0}u_{3t} + w_{1t}$$

is zero. In practice, if the null hypothesis of a zero partial correlation between u_{1t} and u_{3t} cannot be rejected, an exclusion restriction on the corresponding element of B_0 is imposed, and we set $b_{13,0} = 0$.

Similarly, the result of a partial correlation analysis in a four-dimensional VAR system may be

$$\begin{bmatrix} * & 0 & 0 & 0 \\ * & * & 0 & 0 \\ * & 0 & * & 0 \\ 0 & * & * & * \end{bmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ u_{4t} \end{pmatrix} = \begin{pmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{pmatrix},$$

where asterisks denote unrestricted elements. Such a result would suggest that the DGP is compatible with a recursive structure and that there are in fact over-identifying zero restrictions. This information is obviously helpful in identifying mutually uncorrelated shocks.

Of course, there is no guarantee that a fully identified B_0 matrix is obtained using this approach. For example, an outcome such as

$$\begin{bmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ u_{4t} \end{pmatrix} = \begin{pmatrix} w_{1t} \\ w_{2t} \\ w_{3t} \\ w_{4t} \end{pmatrix},$$

where again * denotes elements that may be nonzero, is possible as well. In the latter case, the B_0 matrix is not fully identified, although there are more zero elements than required by the order condition. The problem is that the upper left-hand and lower right-hand submatrices are not identified. Full identification requires additional identifying restrictions from other sources.

Although this data-based approach to identification has some appeal, it also has some obvious drawbacks. First, the fact that the data admit a set of zero elements for B_0 does not mean that the B_0 matrix obtained by the graph-theoretic approach reflects the actual economic structure, as we already illustrated by example in Section 8.2. Put differently, even a finding of uncorrelated reduced-form residuals with partial correlations of zero does not rule out that the variables are related by an instantaneous economic causal structure. In fact, it is not clear how to interpret statements about one variable causing another from an economic point of view, because structural shocks in simultaneous equations models in general are not associated with specific observables.

Second, in practice, the graph-theoretic approach relies on statistical tests to determine whether the partial correlations in the data are zero or not. An obvious concern is that a failure to reject the null hypothesis does not necessarily imply the validity of the null model. Especially in small samples, statistical tests are likely not to reject an invalid null hypothesis of zero partial correlation simply due to low power. Moreover, the full set of zero restrictions is the result of a multiple testing procedure with unknown overall size.

In short, this data-driven identification approach is not well-suited for uncovering economically meaningful structures. This fact helps explain why, as of now, there are not many structural VAR applications based on this approach.

8.6 Summary

The central question in structural VAR analysis is how to recover the unknown parameters of the structural model from the parameters of the reduced-form representation of this model. Knowledge of these structural parameters allows us to characterize the responses of the model variables to the structural shocks and related statistics of interest (see Chapter 4). The problem of solving for the unknown structural parameters as functions of the parameters of the reduced form is known as the identification problem. Solving this identification problem requires the user to impose additional identifying restrictions. As long as we impose enough identifying restrictions, all parameters of the structural model may be uniquely solved for. The choice of these restrictions determines the numerical values of the structural model parameters. It is essential that these identifying restrictions be economically meaningful. Imposing ad hoc restrictions without economic justification will fail to identify correctly the underlying structural model parameters. The challenge for applied users thus is to come up with identifying restrictions that are economically plausible.

In this chapter, we examined the conditions required for the unique identification of the parameters of the structural VAR model from the reduced-form representation. We discussed the role of normalizing assumptions for identification and the challenges in coming up with economically credible identifying restrictions. There are many different types of identifying restrictions, but the most common approach in applied work has been to impose exclusion restrictions on selected elements of the structural impact multiplier matrix (or its inverse). Because such restrictions only affect the contemporaneous interaction among the model variables, conditional on their past values, they are also known as short-run identifying restrictions.

We illustrated the use of short-run identifying restrictions by example. One common assumption has been that there is a causal ordering among the model variables. Such recursively identified structural models are usually difficult to motivate from an economic point of view. Many applications of this approach have been rightly criticized as atheoretical in that the implied shocks are mutually uncorrelated but devoid of economic content and hence not truly structural. If we are willing to restrict attention to the identification of only one structural shock, recursive models may sometimes be given a semistructural interpretation. The latter class of models has played an important role in monetary economics, for example, but is not without serious limitations. A less common alternative approach has been to allow the structural model to be nonrecursive, conditional on past data. Such restrictions can sometimes be derived from an explicit economic model structure.

In response to the scarcity of credible short-run exclusion restrictions in applied work, a large number of alternative (or sometimes complementary) types of restrictions has been developed in the structural VAR literature since the 1990s. Given the widespread use of short-run identifying restrictions in applied work, it is useful to examine in detail the estimation of models identified by short-run restrictions in Chapter 9, before discussing these alternative approaches.