10 Identification by Long-Run Restrictions

Finding enough short-run identifying restrictions can be a challenge in practice. One alternative idea has been to impose restrictions on the long-run response of variables to shocks. In the presence of unit roots in some variables, but not in others, this approach may allow us to identify at least some structural shocks. The promise of this alternative approach to identification is that it allows us to dispense with the controversy about what the right short-run restrictions are and to focus on long-run properties of models that most economists can more easily agree on. For example, it has been observed that most economists agree that demand shocks such as monetary policy shocks are neutral in the long run, whereas productivity shocks are not.

10.1 The Traditional Framework for Imposing Long-Run Restrictions

The idea of imposing long-run restrictions on structural VAR models was first proposed by Blanchard and Quah (1989) in the context of a bivariate model of the U.S. economy. It is useful to review their model for expository purposes. Blanchard and Quah's model attributes variation in U.S. real GDP and unemployment to an aggregate supply shock, w_t^{AS}, and an aggregate demand shock, w_t^{AD}. These structural shocks are identified by imposing that w_t^{AD} has no long-run effect on the level of real GDP. Let ur_t denote the U.S. unemployment rate and gdp_t the log of U.S. real GDP. Define

$$z_t = \begin{pmatrix} \Delta gdp_t \\ ur_t \end{pmatrix} \sim I(0),$$

where by assumption $z_t \sim I(0)$, but $gdp_t \sim I(1)$. The vector z_t is assumed to be generated by a reduced-form VAR process

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

where $A(L) = I_2 - A_1 L - \cdots - A_p L^p -$, such that $A(1) = I_2 - A_1 - \cdots - A_p$, and $u_t \sim (0, \Sigma_u)$ is white noise. The corresponding structural form is

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

where $B(L) = B_0 - B_1 L - \cdots - B_p L^p = B_0 A(L)$ and, hence, $B(1) = B_0 - B_1 - \cdots - B_p = B_0 A(1)$. We impose the normalization $w_t = \left(w_t^{AS}, w_t^{AD}\right)' \sim (0, I_2)$. Given $w_t = B_0 u_t$, it follows that $u_t = B_0^{-1} w_t$ and, hence, $\Sigma_u = B_0^{-1} B_0^{-1\prime}$.

The effect of structural shocks, w_t, on the observed variables is obtained from the structural MA representation

$$z_t = B(L)^{-1} w_t = \Theta(L) w_t.$$

Because z_t is $I(0)$, the effect of any one structural shock on z_t will approach zero as the horizon increases. In other words, both Δgdp_t and ur_t by construction will return to their initial values eventually. This does not mean, however, that the level of real GDP will necessarily return to its initial value. The effect of a given structural shock on gdp_t is the cumulative sum of its effects on Δgdp_t. The long-run cumulative effects are summarized by the matrix $\Theta(1) = \sum_{i=0}^{\infty} \Theta_i = B(1)^{-1}$.

Requiring gdp_t to return to its initial level in the long run in response to an aggregate demand shock imposes an exclusion restriction on the upper right element of $\Theta(1)$ such that

$$\Theta(1) = \begin{bmatrix} \theta_{11}(1) & 0 \\ \theta_{21}(1) & \theta_{22}(1) \end{bmatrix}.$$

In contrast, $\theta_{11}(1)$ remains unrestricted, because aggregate supply shocks affect the level of real GDP in the long run. Moreover, there are no restrictions on the second row of $\Theta(1)$ because the cumulative responses of a stationary variable such as ur_t are clearly different from zero in general.

This example illustrates that if one structural shock is subject to a long-run restriction and the other one is not, it becomes possible to distinguish between these structural shocks. Using the relationship

$$\Theta(1) = B(1)^{-1} = A(1)^{-1} B_0^{-1},$$

it is seen that the exclusion restriction on $\Theta(1)$ is effectively an implicit restriction on B_0 because the reduced-form parameters $A(L)$ are given by the DGP. Thus, long-run restrictions provide an additional source of identifying restrictions for structural VAR models. Moreover, unlike in many of the models we reviewed in Chapter 8, B_0^{-1} (or equivalently B_0) is not restricted to be recursive. Although the choice of real GDP growth as the first variable in z_t is crucial for the interpretation of the Blanchard-Quah model, the choice of the second variable is not. In principle, any other stationary U.S. macroeconomic variable such as the U.S. capacity utilization rate would have done just as well from an econometric point of view.

Given knowledge of the reduced-form VAR model parameters, the unknown elements of $\Theta(1)$ may be recovered from

$$\begin{aligned}\Sigma_u &= B_0^{-1}B_0^{-1\prime}\\ &= \left[A(1)B(1)^{-1}\right]\underbrace{\left[A(1)B(1)^{-1}\right]'}_{\left[B(1)^{-1}\right]'A(1)'}.\end{aligned}$$

Pre-multiplying both sides by $A(1)^{-1}$ and post-multiplying both sides by $[A(1)^{-1}]' = [A(1)']^{-1}$ we obtain

$$\begin{aligned}A(1)^{-1}\Sigma_u[A(1)^{-1}]' &= A(1)^{-1}A(1)B(1)^{-1}\left[B(1)^{-1}\right]'A(1)'\left[A(1)'\right]^{-1}\\ &= \left[B(1)^{-1}\right]\left[B(1)^{-1}\right]'\\ &= \Theta(1)\Theta(1)',\end{aligned}$$

where the left-hand side depends only on reduced-form parameters. Given the symmetry of Σ_u about its main diagonal, we need $K(K-1)/2$ restrictions on $\Theta(1)$, where K is the number of variables in the model, to satisfy the order condition for exact identification of the parameters in $\Theta(1)$. This condition is satisfied by the exclusion restriction imposed by Blanchard and Quah (1989), allowing us to solve for the remaining elements of $\Theta(1)$. If the structure of $\Theta(1)$ is lower triangular, as in the Blanchard-Quah example, this may be accomplished by applying a lower-triangular Cholesky decomposition to $A(1)^{-1}\Sigma_u[A(1)^{-1}]' = \Phi(1)\Sigma_u\Phi(1)'$. Given knowledge of $\Theta(1)$ and $A(1)$, we can recover

$$B_0^{-1} = A(1)\Theta(1),$$

and, once we know B_0^{-1}, we can proceed with the further analysis of the VAR model exactly as in the case of short-run identifying restrictions.

Most applications of long-run restrictions involve a close variation on the theme of Blanchard and Quah (1989), in which the aggregate supply shock is interpreted as an aggregate productivity shock or as a technology shock with permanent effects on real output. Even if more variables are included in VAR models based on long-run restrictions, the focus often is on identifying the responses to aggregate productivity shocks only as opposed to other structural shocks. Such extensions are straightforward. If we augment the original Blanchard and Quah model to include additional stationary variables x_t such that

$$z_t = \begin{pmatrix}\Delta gdp_t\\ ur_t\\ x_t\end{pmatrix} \sim I(0),$$

for example, the aggregate supply shock may be identified by imposing a lower-triangular structure on $\Theta(1)$, as long as we are only interested in identifying the responses to the aggregate supply shock. The higher-dimensional models in Galí (1999) are a good example for this approach. There are other examples of the use of long-run restrictions, however, that involve models that are fully identified, possibly in conjunction with other identifying restrictions (see Section 10.4).

In Section 10.2 we present a formal framework due to King, Plosser, Stock, and Watson (1991) for imposing long-run restrictions on the effects of structural shocks, followed by several empirical examples in Section 10.3. Although our analysis focuses on models in which all variables are $I(0)$ or $I(1)$, as is common in applied work, it should be kept in mind that the idea of long-run restrictions in structural VAR models may be generalized to processes that are integrated of higher order or fractionally integrated (see, e.g., Tschernig, Weber, and Weigand 2013). Our framework is also general enough to accommodate the imposition of additional restrictions on the impact effects of structural shocks. Thus, it can be used to combine long-run and short-run identifying restrictions. For examples, the reader is referred to Section 10.4. In Section 10.5 we review the limitations of long-run identifying restrictions. A more detailed discussion of the estimation of structural VAR models subject to long-run restrictions can be found in Chapter 11.

10.2 A General Framework for Imposing Long-Run Restrictions

Whereas Blanchard and Quah (1989) restricted the cumulative response of the growth rate of real GDP in a stationary VAR representation, a more general framework for studying the long-run effects of structural shocks utilizes the vector error correction representation of the VAR model.

10.2.1 The Long-Run Multiplier Matrix

Suppose that at least some components of the K-dimensional VAR(p) process y_t are $I(1)$. In Chapter 3 we demonstrated that in this situation the VECM,

$$\Delta y_t = \alpha\beta' y_{t-1} + \mathbf{\Gamma}_1 \Delta y_{t-1} + \cdots + \mathbf{\Gamma}_{p-1}\Delta y_{t-p+1} + u_t, \tag{10.2.1}$$

is a convenient reparameterization of the VAR process, where it is assumed that the cointegrating rank is r and that α and β are $K \times r$ matrices of rank r, the former being the loading matrix and the latter the cointegration matrix. The other symbols have their usual meaning. In particular, $u_t \sim (0, \Sigma_u)$ is the reduced-form white noise error term with nonsingular covariance Σ_u. At this point we do not consider deterministic terms because they are not relevant for structural identification.

In Chapter 3 we also showed that this process has the Granger representation

$$y_t = \Xi \sum_{i=1}^{t} u_i + \Xi^*(L)u_t + y_0^*, \tag{10.2.2}$$

where

$$\Xi = \beta_\perp \left[\alpha'_\perp \left(I_K - \sum_{i=1}^{p-1} \mathbf{\Gamma}_i \right) \beta_\perp \right]^{-1} \alpha'_\perp, \tag{10.2.3}$$

$\beta_\perp$ and $\alpha_\perp$ are orthogonal complements of β and α, respectively,[1] $\Xi^*(L)u_t = \sum_{j=0}^{\infty} \Xi_j^* u_{t-j}$ is an $I(0)$ process, and y_0^* contains the initial values. The matrix Ξ has rank $K - r$, i.e., its rank equals the dimension of the process minus the cointegration rank, or, in other words, the rank of Ξ equals the number of common trends.

Since the structural shocks are obtained from the reduced-form errors by a linear transformation, $w_t = B_0 u_t$, we can replace u_t by $B_0^{-1} w_t$ in the Granger representation to obtain

$$y_t = \mathbf{\Upsilon} \sum_{i=1}^{t} w_i + \Xi^*(L)B_0^{-1} w_t + y_0^*, \tag{10.2.4}$$

where $\mathbf{\Upsilon} = [\zeta_{kl}] = \Xi B_0^{-1}$. This representation is useful because it directly shows the long-run effects or permanent effects of the structural shocks on the level of the variables y_t. The matrix $\mathbf{\Upsilon}$ is also known as the matrix of long-run multipliers. Since the coefficient matrices $\Xi_j^* B_0^{-1}$ in the stationary term $\Xi^*(L)B_0^{-1} w_t$ taper off to zero, as $j \to \infty$, it is clear that the long-run effects of the structural shocks can be obtained from $\mathbf{\Upsilon}$. This matrix also has rank $K - r$, just as Ξ, because it is obtained from Ξ by a nonsingular transformation.

Restrictions on the long-run effects of the shocks can be imposed directly on $\mathbf{\Upsilon}$. If a shock does not have any long-run effects at all, the corresponding column in $\mathbf{\Upsilon}$ is restricted to zero. Expressing the Blanchard-Quah example in this notation yields

$$y_t = \begin{pmatrix} gdp_t \\ ur_t \end{pmatrix} = \begin{bmatrix} \zeta_{11} & 0 \\ \zeta_{21} & \zeta_{22} \end{bmatrix} \sum_{i=1}^{t} \begin{pmatrix} w_i^{AS} \\ w_i^{AD} \end{pmatrix} + \dots,$$

where the stationary terms have been suppressed. The $(1, 2)$ element of the $\mathbf{\Upsilon}$ matrix is zero because the aggregate demand shock does not have a long-run effect on gdp_t. Note that we rearranged the original Blanchard-Quah model

[1] Recall that if M is an $m \times n$ matrix of full column rank, an orthogonal complement of M, denoted by $M_\perp$, is an $m \times (m - n)$ matrix with $\mathrm{rk}(M_\perp) = m - n$ such that $M' M_\perp = 0_{n \times (m-n)}$. The orthogonal complement of a nonsingular square matrix is 0 and the orthogonal complement of a zero matrix is I_m.

such that even gdp_t is included in levels in y_t. Whereas y_t contains an $I(1)$ variable, the variable z_t in the introductory example of this chapter is $I(0)$.

Since ur_t is assumed to be $I(0)$, the second row of $\boldsymbol{\Upsilon}$ must be zero as well, because no structural shock can have a nonzero long-run effect on a stationary variable. In other words, we obtain a long-run effects matrix

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \zeta_{11} & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus, the aggregate demand shock has no long-run effects at all. The element ζ_{11}, in contrast, must be nonzero because the rank of $\boldsymbol{\Upsilon}$ is one if only one of the two variables is $I(0)$. In this case the cointegrating rank is one, as shown in Chapter 3, and, thus, $K - r = 1$ for this bivariate system.

More generally, whenever there are $I(0)$ components in y_t, the corresponding row of $\boldsymbol{\Upsilon}$ is zero. In the special case of a stationary VAR process, the cointegrating rank is K and there are no common trends, i.e., $\boldsymbol{\Upsilon} = 0_{K\times K}$. As mentioned earlier, shocks cannot have permanent effects on $I(0)$ variables, so this result makes sense.

In contrast, when all variables are $I(1)$ and there is no cointegration such that $r = 0$, then β and α are zero matrices and their orthogonal complements are simply $K \times K$ identity matrices. Thus,

$$\boldsymbol{\Xi} = \left(I_K - \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \right)^{-1}$$

and

$$\boldsymbol{\Upsilon} = \left(I_K - \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \right)^{-1} B_0^{-1}. \tag{10.2.5}$$

In that case, the first differences of y_t have a reduced-form VAR($p - 1$) representation

$$\Delta y_t = \boldsymbol{\Gamma}_1 \Delta y_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta y_{t-p+1} + u_t$$

and a structural-form representation

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

where $B(L) = B_0 \boldsymbol{\Gamma}(L) = B_0(I_K - \boldsymbol{\Gamma}_1 L - \cdots - \boldsymbol{\Gamma}_{p-1} L^{p-1})$. The structural impulse responses are obtained from the structural MA representation

$$\Delta y_t = B(L)^{-1} w_t = \boldsymbol{\Gamma}(L)^{-1} B_0^{-1} w_t,$$

and the long-run effects on the levels y_t are the cumulated impulse responses $\boldsymbol{\Upsilon} = \boldsymbol{\Gamma}(1)^{-1} B_0^{-1}$ (see expression (10.2.5)).

10.2.2 Identification of Structural Shocks

Section 10.2.1 demonstrated that restrictions on the long-run effects of shocks can be placed directly on the $K \times K$ matrix $\boldsymbol{\Upsilon}$. Thus, we can identify some or potentially all shocks by restrictions on this matrix. For example, if a specific shock is known to have no long-run effect on a particular variable, a zero restriction can be placed on the corresponding element of $\boldsymbol{\Upsilon}$ as in the Blanchard-Quah example.

In imposing restrictions on $\boldsymbol{\Upsilon}$, the properties of this matrix have to be taken into account. In particular, it is important to remember that the matrix has reduced rank $K - r$. An immediate implication of this property is that at most r shocks can have transitory effects only. In other words, at most r columns can be zero. This does not mean, however, that exactly r shocks must be purely transitory. In fact, all shocks can have permanent effects, because a $K \times K$ matrix of rank $K - r$ does not necessarily have zero elements. For example, in a bivariate system, the matrix

$$\boldsymbol{\Upsilon} = \begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix}$$

has rank one, although all elements are nonzero. We first consider the simpler case of $r = 0$, in which $\boldsymbol{\Upsilon}$ has full rank K, before turning to the more difficult case, in which $\boldsymbol{\Upsilon}$ has rank $0 < r < K$.

Unit Roots without Cointegration. One case that is particularly easy to handle arises when the cointegrating rank r is zero such that Δy_t is stationary and, thus, $\boldsymbol{\Upsilon}$ is nonsingular. In that case, identification may be achieved, for example, by specifying a lower-triangular long-run effects matrix. The implied restrictions on B_0 are easy to impose by using the relationship (10.2.5). Defining $\boldsymbol{\Gamma}(L) = I_K - \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i L^i$, that relationship implies

$$\begin{aligned} \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' &= \boldsymbol{\Gamma}(1)^{-1} B_0^{-1} B_0'^{-1} \boldsymbol{\Gamma}(1)'^{-1} \\ &= \boldsymbol{\Gamma}(1)^{-1} \Sigma_u \boldsymbol{\Gamma}(1)'^{-1}. \end{aligned}$$

The latter expression is easy to compute from the reduced form. A lower-triangular $\boldsymbol{\Upsilon}$ can then be obtained from the lower-triangular Cholesky decomposition of $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}'$ as

$$\boldsymbol{\Upsilon} = \operatorname{chol}(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}').$$

Hence,

$$B_0 = \left[\boldsymbol{\Gamma}(1) \operatorname{chol}(\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}') \right]^{-1}.$$

This simple device for computing the restricted B_0 matrix and thereby the structural shocks has proved attractive for applied work. Applications of this approach can be found in Binswanger (2004) and Lütkepohl and Velinov (2016).

Unit Roots with Cointegration. If $0 < r < K$, the analysis becomes more complicated because the rank of $\boldsymbol{\Upsilon}$ must be taken into account when determining the number of restrictions that have to be imposed for full identification of the structural shocks. Recall that in the standard setup $K(K-1)/2$ restrictions have to be imposed on B_0 or on the impact multiplier matrix B_0^{-1} to identify the structural shocks. In Chapter 8 we saw that in that context identification may be achieved by imposing a recursive structure such that B_0 and B_0^{-1} are lower triangular. In contrast, in the present context, simply counting the zero restrictions on the long-run multiplier matrix is not enough to ensure identification. Recursive restrictions on the matrix $\boldsymbol{\Upsilon}$ alone will not achieve identification because of its reduced rank. Consider, for example, the 3×3 matrix

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \zeta_{11} & 0 & 0 \\ \zeta_{21} & \zeta_{22} & 0 \\ \zeta_{31} & \zeta_{32} & \zeta_{33} \end{bmatrix}.$$

Assuming that ζ_{33} is nonzero, this matrix can have rank 1 only if also $\zeta_{11} = \zeta_{21} = \zeta_{22} = 0$. Thus, $\boldsymbol{\Upsilon}$ cannot be lower triangular with all elements below the diagonal nonzero. In fact, imposing the two zero restrictions in the last column implies that the first two elements in columns one and two also must be zero. Hence, they do not count as separate identifying restrictions. Besides, there may be $I(0)$ components in y_t that imply zero rows for $\boldsymbol{\Upsilon}$ and do not serve as identifying restrictions. In other words, simply counting zero restrictions on the long-run multiplier matrix is not enough to ensure identification of the structural shocks. Only if the variables are not cointegrated, and hence $\boldsymbol{\Upsilon}$ has full rank K, is it possible to identify all K shocks by a recursive structure on the long-run effects or, equivalently, by specifying $\boldsymbol{\Upsilon}$ to be triangular.

It is useful to illustrate in more detail the nature of this problem. Recall that, if there is only one shock with purely transitory effects, this shock is identified as the complement to the shocks with permanent effects. If there are two or more shocks with only transitory effects, in contrast, these shocks must be identified by additional restrictions on B_0 or B_0^{-1}. Suppose a $K \times K$ matrix $\boldsymbol{\Upsilon}$ has rank $K - r$ and there are exactly r purely transitory shocks with no long-run effects at all, where $r > 1$. Let the last r shocks be the transitory shocks. Then $\boldsymbol{\Upsilon}$ has the form

$$\boldsymbol{\Upsilon} = [\boldsymbol{\Upsilon}_1, 0_{K\times r}].$$

In other words, the last r columns are zero. Clearly, the last r shocks must be distinguished by features other than their long-run effects. This can be accomplished, for example, by imposing restrictions on the last r columns of the transformation matrix B_0 or its inverse. The first $K - r$ shocks that have permanent effects may be identified by restrictions on $\boldsymbol{\Upsilon}_1$. This discussion highlights that occasionally it may be necessary to complement long-run identifying restrictions with short-run restrictions on B_0 or B_0^{-1}.

It is important to keep in mind, however, that, given the reduced form of the VAR model, all such restrictions jointly constrain the transformation matrix B_0. Suppose that we have a set of linear restrictions

$$R_l \text{vec}(\boldsymbol{\Upsilon}) = r_l \quad \text{or} \quad R_l \text{vec}(\Xi B_0^{-1}) = r_l$$

on the long-run multiplier matrix, where R_l is a suitable given restriction matrix and r_l a given fixed vector. These restrictions can be written as

$$R_l(I_K \otimes \Xi)\text{vec}(B_0^{-1}) = r_l \tag{10.2.6}$$

by using the rules for the vec operator and the Kronecker product. Because Ξ is fully determined by the reduced form (see expression (10.2.3)), it is not constrained by the structural identifying restrictions. In fact, expression (10.2.6) shows that the structural restrictions can be represented as linear restrictions on $\text{vec}(B_0^{-1})$,

$$R_L \text{vec}(B_0^{-1}) = r_l$$

with a restriction matrix $R_L = R_l(I_K \otimes \Xi)$.

If there are additional linear restrictions on the impact effects,

$$R_S \text{vec}(B_0^{-1}) = r_s,$$

then the long-run restrictions and the short-run restrictions can be combined as

$$\begin{bmatrix} R_L \\ R_S \end{bmatrix} \text{vec}(B_0^{-1}) = \begin{pmatrix} r_l \\ r_s \end{pmatrix}. \tag{10.2.7}$$

For a fully identified set of structural shocks there must be at least $K(K-1)/2$ linearly independent restrictions, i.e., the restriction matrix

$$\begin{bmatrix} R_L \\ R_S \end{bmatrix}$$

must have a rank of at least $K(K-1)/2$.

If overidentifying restrictions are considered, there are some further restrictions that have to be taken into account due to the reduced rank of the long-run multiplier matrix $\boldsymbol{\Upsilon}$. As pointed out earlier, in a model with cointegrating rank r, at most r of the structural innovations can have purely transitory effects and at least $K - r$ of them must have permanent effects. Lütkepohl (2008) shows that this fact limits the number of exclusion restrictions we can impose on B_0^{-1}. He proves that, under weak conditions, the number of admissible zero restrictions placed on columns of B_0^{-1} associated with transitory shocks cannot exceed $r - 1$. For example, if $r = 1$ and there is one transitory shock, as in

Blanchard and Quah (1989), there cannot be any zero restriction on the column of B_0^{-1} corresponding to the transitory shock. This result is intuitive, because in the bivariate model, the transitory shock is identified as the residual after identifying the permanent shock, so no further restrictions on B_0^{-1} are required. If $r = 2$ and $K = 3$, there can be at most one zero restriction on each of the columns of B_0^{-1} associated with the transitory shocks. If more zero restrictions are imposed, then Σ_u becomes singular. Such a singularity would contradict the premise that there must be as many shocks as variables in the structural VAR model. The same argument can also be invoked in reverse, by starting with the premise of r transitory shocks, which limits the number of zero restrictions that can be placed on the columns of the long-run structural multiplier matrix that are associated with the permanent shocks.

Although this discussion suggests that it can be difficult to combine short-run and long-run restrictions because the restriction accounting becomes more complicated, it must be kept in mind that overidentifying restrictions are the exception rather than the rule in structural VAR analysis. There are in fact situations in which the additional flexibility offered by long-run restrictions is helpful in achieving full identification. Also, in some cases a specific shock of interest may be particularly easy to identify in the present framework. For example, one may only be interested in a shock with permanent effects in a model where only one such shock is present.

10.3 Examples of Long-Run Restrictions

The approach in Section 10.2.2 requires expressing the VAR model as a VECM or, in the absence of cointegration, as a VAR model in first differences. This section presents a number of examples from the literature.

10.3.1 A Real Business Cycle Model with and without Nominal Variables

King, Plosser, Stock, and Watson (1991) apply the general framework considered in the previous section to the analysis of the Real Business Cycle (RBC) model. Their baseline model includes real output, real consumption, and real investment. An extended model includes in addition the aggregate price level, nominal interest rates, and money holdings.

The Baseline Three-Variable Model. It is useful to start with the baseline VAR model including only logged data for U.S. real output (gnp_t), real consumption (c_t), and real investment (inv_t). Unlike in Blanchard and Quah (1989), in this model all real variables are affected by the same productivity shock in the long run. Given a productivity shock with a stochastic trend, balanced growth under uncertainty implies that real consumption, real investment, and real output are cointegrated such that $c_t - gnp_t \sim I(0)$ and $inv_t - gnp_t \sim I(0)$. This means that the VAR model for $y_t = (gnp_t, c_t, inv_t)'$ may equivalently be written as a reduced-form VECM as in the previous section with cointegration rank $r = 2$ and known cointegrating matrix β.

King et al. are interested in using this model to identify the responses to the common productivity shock. The remaining two transitory shocks remain economically unidentified. In this sense, the model is only partially identified. Assuming that there are indeed two transitory shocks and placing them last in the vector of shocks such that the first shock is the permanent shock, we obtain

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \end{bmatrix},$$

where $*$ denotes an unrestricted element. No further restrictions are necessary to identify the permanent shock.

It is possible to identify the remaining two transitory shocks in this model by combining the restrictions on the long-run multiplier matrix with short-run identifying restrictions. For local just-identification of all structural elements of B_0^{-1}, we need $K(K-1)/2 = 3$ restrictions in this model. Since the long-run effects matrix $\boldsymbol{\Upsilon}$ has rank $K - r = 1$, the two zero columns stand for two independent restrictions only. Clearly, the transitory shocks are not identified without further restrictions. One restriction on the last two columns of B_0^{-1} is sufficient to disentangle the two transitory shocks. For example, we may impose

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \end{bmatrix} \quad \text{and} \quad B_0^{-1} = \begin{bmatrix} * & * & * \\ * & * & 0 \\ * & * & * \end{bmatrix},$$

where $*$ indicates again that no restriction is imposed, which allows the first transitory shock (w_{2t}) to have instantaneous effects on all variables and prevents the second transitory shock (w_{3t}) from having an impact effect on real consumption. Whether such a restriction makes economic sense is a different matter. Indeed, in the baseline model of King et al. one would be hard-pressed to justify an additional exclusion restriction on B_0^{-1}, and the authors are content to focus on the effects of the balanced growth shock, w_{1t}. Which restriction is used to statistically identify the transitory shocks leaves the permanent shock and its effects unaffected. This means that we may simply impose an arbitrary exclusion restriction on the last two columns of B_0^{-1} if all we are interested in is the responses to the common productivity shock.

King, Plosser, Stock, and Watson (1991) investigate the ability of w_{1t} to explain the variability of the three model variables. They find impulse response patterns that are consistent with simple theoretical models in that all three variables increase in response to a positive balanced growth shock, but real output and real investment respond more strongly than real consumption. Most of the adjustment is complete within four years. Structural forecast error variance decompositions indicate that 45–58% of the variability of real GNP growth at the short horizons is explained by the balanced-growth shock. This increases to 68% at the two-year horizon and 81% at the six-year horizon.

An equivalent way of writing the King et al. model is as a stationary VAR model for $z_t \equiv (\Delta gnp_t, c_t - gnp_t, inv_t - gnp_t)'$, where the balanced growth shock affects gnp_t in the long run, but not the stationary ratios $c_t - gnp_t$ and $inv_t - gnp_t$. This is, in essence, the representation chosen by Blanchard and Quah (1989). If we dropped the last variable in z_t, the VAR model for $(\Delta gnp_t, c_t - gnp_t)'$ could be analyzed using exactly the same approach used by Blanchard and Quah for $(\Delta gdp_t, ur_t)'$. In this sense, the analysis in King et al. may be viewed as a generalization of the approach in Blanchard and Quah (1989).

As mentioned earlier, the approach of relying on the stationary VAR representation rather than VECMs is also common when working with larger models. For example, Galí (1999) fits a stationary VAR model to $z_t \equiv (\Delta prod_t, \Delta h_t, \Delta m_t - \Delta p_t, i_t - \Delta p_t, \Delta^2 p_t)'$, where $prod_t$ denotes labor productivity, h_t stands for hours worked, m_t denotes money holdings, p_t the aggregate price level, and i_t the nominal interest rate. His identifying assumption is that only technology shocks have long-run effects on labor productivity. The four non-technology shocks in the model are not individually identified from an economic point of view. Thus, they can be identified arbitrarily from a statistical point of view. For example, one may restrict the accumulated effects matrix $\Theta(1)$ to be lower triangular. In this setup, as long as we are only interested in identifying the first structural shock, the ordering of the other shocks is inconsequential because the structural responses to the first shock will be invariant to the ordering of the identifying assumptions for the remaining variables.

In the baseline three-variable model of King, Plosser, Stock, and Watson (1991) this approach would involve specifying a VAR model for $z_t = (\Delta gnp_t, c_t - gnp_t, inv_t - gnp_t)' \sim I(0)$ and imposing

$$\Theta(1) = \begin{bmatrix} * & 0 & 0 \\ * & * & 0 \\ * & * & * \end{bmatrix}.$$

The key difference is that in this case no further restrictions on B_0^{-1} are required because the restrictions on $\Theta(1)$ suffice to pin down the impact responses to the common productivity shock. However, the response of c_t to a productivity shock can only be constructed by cumulating the response of Δgnp_t and adding the implied response of gnp_t to that of $c_t - gnp_t$ (and similarly for inv_t). This discussion illustrates that the way we impose the identifying assumptions implied by a given economic model in estimation depends on how the structural VAR model is specified.

The Extended Six-Variable Model with Nominal Variables. The analysis becomes substantially more complicated once we allow for the possibility that there is more than one permanent shock. This situation is illustrated by the second example considered in King, Plosser, Stock, and Watson (1991). King et al. extend the baseline model such that $y_t = (gnp_t, c_t, inv_t, m_t - p_t, i_t, \Delta p_{t+1})'$, where i_t is the nominal interest rate, p_t is the log of the price level, and m_t the log of nominal money holdings. In this six-variable VAR system an additional cointegrating relationship arises that represents the money market equilibrium:

$$m_t - p_t - \beta_1 gnp_t + \beta_2 i_t \sim I(0), \tag{10.3.1}$$

where real balances $(m_t - p_t)$, real output, and the nominal interest rate are assumed to be $I(1)$. At the same time, the balanced-growth paths must be allowed to depend on the real interest rate in recognition of the fact that growth theory predicts that higher real interest rates lower the share of output entering investment, while raising the share of output in consumption:

$$c_t - gnp_t - \phi_1(i_t - \Delta p_{t+1}) \sim I(0),$$

$$inv_t - gnp_t - \phi_2(i_t - \Delta p_{t+1}) \sim I(0).$$

Unlike in the baseline three-variable model, the consumption and investment shares are treated as $I(1)$ variables, as is the real interest rate.

Combining these results, we have three cointegrating relationships (and thus three common trends and three permanent shocks) in the model. The first permanent shock is the balanced-growth shock with long-run effects on real balances as well as real output, consumption, and investment; the second permanent shock is an inflation shock that affects the inflation rate and the nominal interest rate in the long run, but has no long-run effects on real output, consumption, or investment; and the third permanent shock is a real interest rate shock with long-run effects on the two ratios, the nominal interest rate, and real balances. These permanent shocks are assumed to be mutually uncorrelated as well as uncorrelated with the transitory shocks. As before, no attempt is made to identify the transitory shocks from an economic point of view.

Assuming that the three transitory shocks are placed last in the vector of shocks, the 6×6 long-run structural multiplier matrix takes the form $\boldsymbol{\Upsilon} = [\boldsymbol{\Upsilon}_1, 0_{6\times 3}]$, where $\boldsymbol{\Upsilon}_1$ is restricted as

$$\underset{(6\times 3)}{\boldsymbol{\Upsilon}_1} = \begin{bmatrix} * & 0 & 0 \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & 0 \end{bmatrix}.$$

Taking into account the results of a detailed cointegration analysis, King et al. decompose this matrix and impose further restrictions compatible with the rank of this matrix. They interpret the three permanent shocks as a balanced-growth shock (w_t^{growth}), a neutral inflation shock ($w_t^{\text{inflation}}$), and a real interest rate shock ($w_t^{\text{real interest}}$). The interpretation of the latter two shocks is not directly motivated based on economic models, and indeed these shocks are not structural in the sense discussed in Chapter 8. If only the balanced-growth shock is of interest, the identification of the other permanent shocks is not important, of course. In other words, the impulse responses to the first shock are not affected by the identifying restrictions for the second and third shocks.

Like in the three-variable model, constructing an estimate of the first column of B_0^{-1} requires additional ad hoc restrictions on the elements of the last three columns of B_0^{-1}. Given that we have already imposed three restrictions on $\boldsymbol{\Upsilon}$ to identify $K - r$ shocks with permanent effects, we need to impose as many additional restrictions on the structural impact multiplier matrix as are required to identify the remaining $r = 3$ structural shocks. This may be accomplished, for example, by setting

$$B_0^{-1} = \begin{bmatrix} * & * & * & * & 0 & 0 \\ * & * & * & * & * & 0 \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \\ * & * & * & * & * & * \end{bmatrix}.$$

King et al. show that in the extended model the explanatory power of the balanced-growth shock for real output is substantially reduced. Much of the short-run variability in output and investment is associated with the permanent real interest rate shock. The permanent inflation shock explains little of the variation in the real variables.

10.3.2 A Model of Neutral and Investment-Specific Technology Shocks

Most models based on long-run restrictions identify only one permanent shock. Fisher (2006) considers a growth model with two permanent shocks. The motivation is that, from a theoretical point of view, conventional technology shocks that are neutral in that they affect the production of all goods homogeneously are not the only possible source of permanent effects on labor productivity. There also are investment-specific technology shocks that are embodied in capital. Omitting the latter shocks from the empirical analysis is likely to bias the VAR estimates of the response to neutral technology shocks. Hence, Fisher designs a model that incorporates both types of technology shocks.

Let $y_t = (p_t, prod_t, h_t)'$, where p_t is the log real price of investment goods, $prod_t$ is the log of labor productivity, and h_t is the log of per capita hours worked. The first two variables are treated as $I(1)$, whereas $h_t \sim I(0)$. Moreover, it is assumed that p_t and $prod_t$ are not cointegrated. Thus, $(\Delta p_t, \Delta prod_t, h_t)'$ is a stationary vector. The cointegrating rank of the model for y_t is $r = 1$ due to the inclusion of one $I(0)$ variable, so there must be at least two shocks with permanent effects.

The three structural shocks are a capital-embodied technology shock, w_t^{cet}, a labor productivity shock, w_t^{lp}, and a transitory shock, w_t^{trans}. They are ordered as $w_t = (w_t^{cet}, w_t^{lp}, w_t^{trans})'$. Clearly the last shock is identified by a zero column in the long-run effects matrix $\boldsymbol{\Upsilon}$. There are three assumptions for identifying the permanent shocks, which are explicitly derived from a real business cycle model. First, only capital-embodied technology shocks have a long-run effect on the log-level of the price of investment goods. Second, both neutral technology shocks and capital-embodied technology shocks have a long-run effect on the log-level of labor productivity. Third, shocks to investment-specific technology raise labor productivity in the long run by an amount that is proportionate to the amount by which they lower the log-level price of investment goods in the long run. The constant of proportionality is presumed known. The third assumption is not necessary for just-identification of the model, but serves as an overidentifying assumption.

These identifying restrictions can be imposed on the matrix of accumulated long-run effects $\Theta(1)$ for the $I(0)$ VAR model for $(\Delta p_t, \Delta prod_t, h_t)'$ as in Fisher (2006). This approach results in

$$\Theta(1) = \begin{bmatrix} * & 0 & 0 \\ * & * & 0 \\ * & * & * \end{bmatrix}$$

with $\theta_{22}(1) = \alpha\theta_{21}(1)$, where $\theta_{ij}(1)$ denotes the ij^{th} element of $\Theta(1)$ and α is known. Fisher (2006) imposes $\alpha = 1/3$. Fisher shows that the results are largely unaffected by the imposition of the overidentifying restriction. Alternatively, and equivalently, these restrictions can be expressed in terms of the structural long-run effects matrix:

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \zeta_{11} & 0 & 0 \\ \zeta_{21} & \alpha\zeta_{21} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where the last row of zeros is due to h_t being $I(0)$ and, hence, none of the shocks moves hours permanently. Fisher's model may be augmented to include additional variables (see Fisher 2006).

10.3.3 A Model of Real and Nominal Exchange Rate Shocks

Although most applications of long-run restrictions focus on productivity shocks, there are other applications. For example, Enders and Lee (1997) propose a bivariate model of real and nominal exchange rates. They distinguish between real shocks that affect both real and nominal exchange rates equally in the long run and nominal shocks that affect only nominal exchange rates in the long run. Based on U.S. dollar exchange rates for Canada, Japan, and Germany since 1973, Enders and Lee find that the real shock explains much of the observed variability of real and nominal exchange rate movements. Their approach is formally identical to that in Blanchard and Quah (1989).

Let r_t denote the real exchange rate between the United States and a foreign country and e_t the corresponding nominal rate. Assuming that both variables are $I(1)$ but not cointegrated, we can set up a VAR model in first differences of the original variables and specify $z_t = (\Delta r_t, \Delta e_t)'$. The structural shocks are $w_t = \left(w_t^{\text{real}}, w_t^{\text{nominal}}\right)'$. As before, we write

$$B_0^{-1} = A(1)\Theta(1),$$

where $\Theta(1)$ is the matrix of accumulated long-run effects of the model in first differences and identification is achieved by restricting the upper right element of $\Theta(1)$ to zero, consistent with the notion of the long-run neutrality of nominal shocks. This restriction suffices to identify both shocks. The precise type of the nominal and real shocks is left unspecified in the model. Enders and Lee (1997), however, make the case that the distinction between nominal and real shocks is consistent with a wide class of theoretical models including the well-known Dornbusch (1976) overshooting model.

An alternative way of imposing this identifying restriction on the matrix of accumulated long-run effects, $\Theta(1)$, would be to work with the framework of Section 10.2 for the levels variables $y_t = (r_t, e_t)'$. The matrix $\Theta(1)$ is identical to the structural long-run effects matrix $\boldsymbol{\Upsilon}$ for the levels variables. Thus, using a VAR model for the levels variables $y_t = (r_t, e_t)'$, the structural shocks w_t are identified by one zero restriction on $\boldsymbol{\Upsilon}$,

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}.$$

10.3.4 A Model of Expectations about Future Productivity

An influential study by Beaudry and Portier (2006) focuses on the problem of capturing shifts in expectations about future productivity. They start with a bivariate model. Let $y_t = (tfp_t, sp_t)'$, where tfp_t denotes the log of quarterly total factor productivity and sp_t the log of the Standard & Poors 500 composite stock price index deflated by the quarterly GDP price deflator. Given that these two variables appear cointegrated, Beaudry and Portier focus on the VECM representation of the VAR model for y_t.

They consider two identification schemes. In the first specification, they impose a recursive ordering on B_0^{-1} such that the second structural shock does not contemporaneously affect tfp_t, while both structural shocks are allowed to affect sp_t instantaneously. Put differently,

$$\begin{pmatrix} u_t^{tfp} \\ u_t^{sp} \end{pmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix}.$$

This identifying assumption is consistent with the view that stock prices incorporate new information about productivity instantaneously and that stock prices anticipate increases in productivity that are yet to come. Beaudry and Portier interpret w_{2t} as a "news shock" by which they mean an anticipated change in future productivity.[2] In contrast, they interpret w_{1t} as an unanticipated productivity shock.

In the second specification, they impose instead a long-run identifying assumption. If we let $y_t = (tfp_t, sp_t)'$, this involves restricting the long-run multiplier matrix as

$$\Upsilon = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix}.$$

In this alternative specification w_{1t} refers to a shock with long-run effects on tfp_t, whereas w_{2t} has no long-run effects on tfp_t. Equivalently, one could work with $z_t = (\Delta tfp_t, sp_t - tfp_t)'$ by analogy to Blanchard and Quah (1989).

Beaudry and Portier observe that the responses to w_{2t} in the first specification and to w_{1t} in the second specification appear very similar and that the respective shock series are almost perfectly correlated. They proceed to show that this finding is robust to augmenting the VAR model to include real consumption or real consumption and hours worked. Beaudry and Portier (2006) conclude from this evidence that these shocks are effectively the same shock, which implies that permanent changes in productivity growth are preceded by stock market booms. They also show that the observed impulse response patterns are qualitatively consistent with theoretical models in which technological innovation affects productivity with a delay.

This result has attracted considerable attention in the business cycle literature because it differs sharply from the conventional view of the business cycle being driven by unanticipated changes in total factor productivity. For example, it explains how booms and busts can happen absent large changes in fundamentals and why no technological regress is required to generate recessions.

[2] In Chapter 7 we noted that this "news shock" terminology is misleading and unrelated to the earlier literature on news shocks properly defined.

It also provides an alternative explanation for the observed comovement of macroeconomic aggregates. Indeed, estimates of VECMs including additional macroeconomic variables suggest that the "news shocks" identified by Beaudry and Portier are associated with increases in consumption, investment, output, and hours on impact and appear to constitute an important source of business cycle fluctuations. These findings have spurred the development of theoretical models capable of generating news-driven comovement among macroeconomic aggregates.

At the same time, there has been growing skepticism about the empirical approach used by Beaudry and Portier (2006). In particular, Kurmann and Mertens (2014) show that in the VECMs with more than two variables estimated by Beaudry and Portier, their identification scheme fails to determine news about total factor productivity. This point is important because the higher-dimensional VECMs are what allows Beaudry and Portier to quantify the business cycle effects of the "news shock". Without that evidence the importance of these shocks remains unclear.

In the words of Kurmann and Mertens (2014), the identification problem arises from the interplay of two assumptions. First, Beaudry and Portier's identification scheme imposes the restriction that one of the non-news shocks has no permanent impact on either TFP or consumption. Second, the VECMs estimated by Beaudry and Portier (2006) postulate that total factor productivity and consumption are cointegrated. As a result, total factor productivity and consumption have the same permanent component, which makes one of the two long-run restrictions redundant and leaves an infinite number of possible solutions with very different implications for the business cycle. The results reported in Beaudry and Portier (2006) represent just one arbitrary choice among these solutions, making it impossible to draw any conclusions about the role of news shocks.

More formally, consider the example of a VECM for

$$y_t = \begin{pmatrix} tfp_t \\ sp_t \\ c_t \end{pmatrix},$$

where all variables are in logs and c_t denotes consumption. Similar to their baseline model, Beaudry and Portier impose the short-run exclusion restriction that news shocks have no contemporaneous effect on tfp_t such that the (1,2) element of B_0^{-1} is zero:

$$B_0^{-1} = \begin{bmatrix} * & 0 & * \\ * & * & * \\ * & * & * \end{bmatrix}.$$

In addition, Beaudry and Portier impose two long-run exclusion restrictions. Note that all three variables share a common trend such that $K = 3$, $r = 2$, and $K - r = 1$. This means that the $K \times K$ long-run multiplier matrix $\boldsymbol{\Upsilon}$ has rank $K - r = 1$. Beaudry and Portier (2006) impose the exclusion restriction that the third structural shock does not effect the level of tfp_t or the level of c_t in the long run such that the $(1, 3)$ and $(3, 3)$ element of $\boldsymbol{\Upsilon}$ are zero. Given that $\boldsymbol{\Upsilon}$ is of rank one, which means that all its rows are proportionate or all its columns are proportionate, we can conclude that either

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & * & 0 \\ * & * & 0 \\ * & * & 0 \end{bmatrix} \quad \text{or} \quad \boldsymbol{\Upsilon} = \begin{bmatrix} 0 & 0 & 0 \\ * & * & * \\ 0 & 0 & 0 \end{bmatrix},$$

where * denotes an unrestricted element. Both specifications are consistent with the identifying assumptions of Beaudry and Portier (2006). Although for the first of these alternative specifications of $\boldsymbol{\Upsilon}$ all three shocks would be identified, this is not the case for the second specification. Given that the long-run restrictions on $\boldsymbol{\Upsilon}$ in the latter case do not contribute to the identification of the structural shocks and given that there is only one restriction imposed on B_0^{-1}, the matrix B_0^{-1} is unidentified. Hence, without further identifying restrictions, the empirical analysis in Beaudry and Portier (2006) is uninformative.

As discussed in Kurmann and Mertens (2014), this problem cannot be addressed by simply not imposing the cointegrating restrictions in estimation, because this change in the model specification does not make the cointegration between tfp_t and c_t disappear, if such cointegration indeed exists in the data. If we imposed the absence of cointegration between tfp_t and c_t, however, disregarding the possible presence of cointegration, the shock implied by the remaining identifying restrictions would be largely unrelated to total factor productivity. In short, the approach employed by Beaudry and Portier (2006) is not informative about the question of interest.

10.4 Examples of Models Combining Long-Run and Short-Run Zero Restrictions

As the previous empirical example illustrated, in models with more than two variables it is common to combine long-run restrictions with short-run zero restrictions on B_0^{-1} or B_0, allowing for additional shocks with transitory effects to be identified. This section provides some additional examples.

10.4.1 The IS-LM Model Revisited

A case in point is the IS-LM model of Galí (1992). Galí's objective is to further disentangle the effects of transitory money demand shocks, money supply shocks, and shocks to the IS curve. His approach is to treat the textbook IS-LM model as a description of the interactions of the VAR model variables conditional on past data. Galí considers a quarterly model for $z_t = (\Delta gnp_t, \Delta i_t, i_t - \Delta p_t, \Delta m_t - \Delta p_t)'$. Movements in these macroeconomic variables are determined by four types of exogenous disturbances: aggregate supply shocks (w_t^{AS}), money supply shocks (w_t^{MS}), money demand shocks (w_t^{MD}), and shocks to the IS curve (w_t^{IS}). Thus, $w_t = (w_t^{AS}, w_t^{MS}, w_t^{MD}, w_t^{IS})'$. Ignoring the lagged dependent variables for expository purposes, the unrestricted structural VAR model can be written as

$$\begin{aligned}
\Delta gnp_t &= -b_{12,0}\Delta i_t - b_{13,0}(i_t - \Delta p_t) - b_{14,0}(\Delta m_t - \Delta p_t) + w_t^{AS},\\
\Delta i_t &= -b_{21,0}\Delta gnp_t - b_{23,0}(i_t - \Delta p_t) - b_{24,0}(\Delta m_t - \Delta p_t) + w_t^{MS},\\
i_t - \Delta p_t &= -b_{31,0}\Delta gnp_t - b_{32,0}\Delta i_t - b_{34,0}(\Delta m_t - \Delta p_t) + w_t^{MD},\\
\Delta m_t - \Delta p_t &= -b_{41,0}\Delta gnp_t - b_{42,0}\Delta i_t - b_{43,0}(i_t - \Delta p_t) + w_t^{IS},
\end{aligned}$$

where $b_{ij,0}$ denotes the ij^{th} element of B_0. Galí interprets the first equation as an aggregate supply function, the second equation as a money supply function, the third equation as the money demand function, and the last equation as an IS function. He imposes six identifying restrictions on the accumulated long-run effects of selected shocks and B_0^{-1}. First, money supply shocks (w_t^{MS}), money demand shocks (w_t^{MD}), and IS shocks (w_t^{IS}) have no long-run effects on real GNP. Only aggregate supply shocks (w_t^{AS}) affect real GNP in the long run. This implies the restrictions $\theta_{12}(1) = \theta_{13}(1) = \theta_{14}(1) = 0$. Second, money demand shocks and money supply shocks do not have contemporaneous effects on output, which distinguishes them from IS shocks. This implies $b_0^{12} = 0$ and $b_0^{13} = 0$. Third, the monetary authority is assumed not to react contemporaneously to changes in the price level. This implies that $b_{23,0} + b_{24,0} = 0$, which imposes a linear restriction on B_0. The restricted cumulated long-run effects $\Theta(1)$ are of the form

$$\Theta(1) = \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}.$$

Equivalently let $y_t = (gnp_t, i_t, i_t - \Delta p_t, m_t - p_t)'$. Then, using the notation of Section 10.2, Galí's restrictions can be expressed in terms of the following constraints on $\boldsymbol{\Upsilon}$:

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & * & * \\ 0 & 0 & 0 & 0 \\ * & * & * & * \end{bmatrix},$$

where * indicates elements that are not explicitly restricted. The third row of zeros in $\boldsymbol{\Upsilon}$ is due to the stationarity of the real interest rate.[3]

Having estimated the structural model, Galí examines how well the model matches traditional Keynesian views. While the timing and magnitude of the structural impulse responses is largely consistent with the predictions of more elaborate New Keynesian models, a structural forecast error variance decomposition suggests that aggregate supply shocks have played a larger role in explaining economic fluctuations than traditional Keynesian views suggest. Historical decompositions indicate that recessions historically were caused by the coincidence of several adverse structural shocks of different types, with the mix of the adverse shocks varying considerably across recessions.

10.4.2 A Model of the Neoclassical Synthesis

A second example is Shapiro and Watson (1988). This study proposes a model of the U.S. economy that exploits insights from neoclassical economics about long-run behavior, while allowing for Keynesian explanations of short-run behavior. Unlike the preceding example, Shapiro and Watson do not take a stand on the economic model underlying the short-run behavior. Let h_t denote the log of hours worked, o_t the price of oil, gdp_t the log of real GDP, π_t inflation, and i_t the nominal interest rate. Shapiro and Watson decompose fluctuations in the $I(0)$ vector $z_t = (\Delta h_t, \Delta o_t, \Delta gdp_t, \Delta \pi_t, i_t - \pi_t)'$ in terms of labor supply shocks (w_t^{LS}), oil price shocks ($w_t^{\text{oil price}}$), technology shocks ($w_t^{\text{technology}}$), and two aggregate demand shocks ($w_t^{\mathrm{AD-IS}}$ and $w_t^{\mathrm{AD-LM}}$). The first identifying assumption is that aggregate demand shocks have no long-run effects on real GDP or hours worked. The second identifying assumption is that the long-run labor supply is exogenous, which allows Shapiro and Watson to separate the effects of shocks to technology from those to labor supply. The third identifying assumption is that exogenous oil price shocks have a permanent effect on the level of all $I(1)$ variables except hours worked. The two aggregate demand shocks may be interpreted as goods market (IS) and money market (LM) shocks. No effort is made to identify the two aggregate demand shocks separately.

Using our framework to state the restrictions formally, we use a vector of levels variables $y_t = (h_t, o_t, gdp_t, \pi_t, i_t - \pi_t)'$. All variables but the real interest rate are assumed to be $I(1)$ and not cointegrated, whereas $i_t - \pi_t \sim I(0)$. With this set of variables and the five shocks $w_t = (w_t^{\mathrm{LS}}, w_t^{\text{oil price}}, w_t^{\text{technology}}, w_t^{\mathrm{AD-IS}}, w_t^{\mathrm{AD-LM}})'$ characterized earlier, we obtain the following matrix of

[3] A critical discussion of the identifying assumptions imposed in this model is provided in Pagan and Pesaran (2008), who show that two out of the three short-run restrictions imposed by Galí (1992) are not required when restrictions consistent with the cointegration properties of the variables are imposed.

long-run effects and B_0:

$$\boldsymbol{\Upsilon} = \begin{bmatrix} * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B_0 = \begin{bmatrix} * & * & * & * & * \\ 0 & * & 0 & 0 & 0 \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix},$$

where * indicates that no explicit restriction is imposed. The last row of zeros in $\boldsymbol{\Upsilon}$ is due to $i_t - \pi_t$ being $I(0)$. Note that the structure of $\boldsymbol{\Upsilon}$ is not recursive. The additional short-run restrictions arise because the change in the price of oil is treated as exogenous white noise.

10.4.3 A U.S. Macroeconomic Model

Fisher, Huh, and Pagan (2016) stress the need for extending the traditional classification of shocks to include shocks with permanent effects on the level of at least one $I(1)$ variable and shocks with transitory effects on all variables. When there are additional $I(0)$ variables included in the VAR model, the corresponding additional shocks may have purely transitory effects on all variables or may have transitory effects on some variables and permanent effects on others. In the latter case, there will be more structural shocks with permanent effects on at least one of the $I(1)$ variables than suggested by the rank of $\boldsymbol{\Upsilon}$. The concern is that in this situation extra care is required to avoid some shocks in the model having unintended permanent effects. For example, nominal shocks may have unintended long-run effects on real variables and relative prices, unless the researcher is careful in specifying the structural model.

This point may be illustrated using the example of Peersman (2005). Peersman postulates a quarterly 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_t), real output growth (Δq_t), consumer price inflation (Δp_t), and the short-term nominal interest rate (i_t). The nominal interest is considered $I(0)$. The three $I(1)$ variables o_t, q_t, and p_t are expressed in first differences. Cointegration among the variables in levels is ruled out. In particular, the real price of oil is implicitly assumed to be $I(1)$. Because we want to study the model within the framework of Section 10.2, we consider the vector of levels variables $y_t = (o_t, q_t, p_t, i_t)$.

The vector of structural shocks includes a nominal oil price shock ($w_t^{\text{oil price}}$), a domestic aggregate supply shock (w_t^{AS}), a domestic aggregate demand shock (w_t^{AD}), and a domestic monetary policy shock ($w_t^{\text{monetary policy}}$). Identification involves two long-run exclusion restrictions and four contemporaneous exclusion restrictions. Neither aggregate demand shocks nor monetary policy shocks are allowed to affect the level of real output in the long run. Accounting also for the stationarity of the nominal interest rate (i_t), the structural long-run effects matrix has the form

$$\Upsilon = \begin{bmatrix} * & * & * & * \\ * & * & 0 & 0 \\ * & * & * & * \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{10.4.1}$$

The short-run restrictions arise from treating the oil price as predetermined with respect to all other variables, providing three exclusion restrictions, and from assuming monetary policy shocks not to have a contemporaneous effect on real output:

$$u_t = \begin{pmatrix} u_t^o \\ u_t^q \\ u_t^p \\ u_t^i \end{pmatrix} = B_0^{-1} w_t = \begin{bmatrix} * & 0 & 0 & 0 \\ * & * & * & 0 \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{pmatrix} w_t^{\text{oil price}} \\ w_t^{\text{AS}} \\ w_t^{\text{AD}} \\ w_t^{\text{monetary policy}} \end{pmatrix}.$$

Writing the long-run restrictions as in equation (10.4.1) reveals an obvious problem with this VAR model specification, which was first highlighted by Fisher, Huh, and Pagan (2016). In particular, the monetary policy shock in this model may have a permanent effect on the real price of oil because the nominal oil price may fall by more than the price level in the long-run, which is inconsistent with the maintained notion of long-run monetary neutrality. Likewise, aggregate demand shocks may have a long-run effect on the level of the real price of oil, invalidating the analysis in Peersman (2005).

To arrive at a VAR model with more economically defensible long-run properties, we may replace the nominal price of oil (o_t) by the real price of oil ($o_t - p_t$), allowing us to impose the required additional long-run restrictions. Specifically, we need to impose that neither aggregate demand nor monetary policy shocks affect the level of the real price of oil and the level of real output in the long run. In addition to these four long-run exclusion restrictions, we require two contemporaneous restrictions for exact identification, one to separate the aggregate demand shock from the monetary policy shock and the other to separate the real oil price shock from the aggregate supply shock. Fisher et al. impose the restrictions that U.S. aggregate demand and U.S. aggregate supply shocks have no contemporaneous effects on the real price of oil. In other words, in a model for $y_t = (o_t - p_t, q_t, p_t, i_t)$ their restrictions for the long-run effects and the impact effects are

$$\Upsilon = \begin{bmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & * \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B_0^{-1} = \begin{bmatrix} * & 0 & 0 & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}.$$

Fisher, Huh, and Pagan (2016) show that price and output puzzles that were absent in the original specification of Peersman (2005) reemerge when the model is restricted to enforce the absence of unintended long-run effects.

10.5 Limitations of Long-Run Restrictions

There are a number of concerns related to using long-run restrictions for identifying structural VAR models. Our discussion in this section focuses on concerns with long-run restrictions that arise within the framework described in Section 10.2. Some of these concerns are of more general nature and arise in one form or another also in the context of short-run restrictions. Others are specific to the use of long-run identifying restrictions. Concerns specifically related to estimating structural VAR models with long-run restrictions that arise from the imprecision of estimates of the long-run impulse responses are addressed in Chapter 11, which focuses on the estimation of structural VAR models based on long-run restrictions.

10.5.1 Long-Run Restrictions Require Exact Unit Roots

One important limitation of the long-run identification schemes presented so far is that they require us to take a stand on the presence of exact unit roots in the autoregressive lag polynomial $A(L)$. This means that this alternative approach is more limited in scope than VAR models based on short-run restrictions, which remain valid regardless of the order of integration of the variables.

One alternative that allows for departures from the exact $I(1)$ hypothesis is to impose long-run identifying restrictions in structural VAR models of fractionally integrated variables (see Chapter 2). Tschernig, Weber, and Weigand (2013) propose an extension of the Granger representation for fractionally integrated variables of the form

$$y_t = \begin{bmatrix} \Delta_+^{-\delta_1} & & 0 \\ & \ddots & \\ 0 & & \Delta_+^{-\delta_K} \end{bmatrix} \Xi B_0^{-1} \sum_{i=1}^{t} w_i + \begin{bmatrix} \Delta_+^{b-\delta_1} & & 0 \\ & \ddots & \\ 0 & & \Delta_+^{b-\delta_K} \end{bmatrix} \Xi_+^*(L_b) B_0^{-1} w_t + y_{0t}^*,$$

where y_{0t}^* denotes the initial values, $\delta_1, \dots, \delta_K$ and b are real numbers, Δ^d stands for the fractional differencing operator defined as $\Delta^d = (1-L)^d \equiv \sum_{i=0}^{\infty}(-1)^i\binom{d}{i}L^i$, and Δ_+^d denotes a truncated version of this expansion. Similarly, $\Xi_+^*(L_b)$ is a truncated operator. The operator L_b denotes the fractional lag operator defined as $L_b \equiv 1 - \Delta^b$. The important point here is that, for certain values of δ_j, restrictions may be placed on the long-run effects of the structural shocks on the model variables by restricting the matrix ΞB_0^{-1}.

Tschernig et al. investigate a bivariate system of U.S. log real GDP (gdp_t) and the log of its implicit deflator (p_t). They specify two shocks that can be interpreted as aggregate demand and aggregate supply shocks. The aggregate demand shock is identified as a shock having no persistent effect on gdp_t. This identifying assumption is imposed by restricting the upper right-hand element of ΞB_0^{-1} to zero. There are no restrictions on the long-run effects of the aggregate supply shock. The latter shock has persistent but not necessarily permanent effects on gdp_t.

While this approach provides an alternative to the exact unit root framework, it also involves additional complications. For example, it requires the user to assess the fractional and cofractional properties of the model variables. In addition, justifying long-run restrictions on economic grounds is likely to be more difficult in this framework than in the VECM framework.

10.5.2 Sensitivity to Omitted Variables

In low-dimensional models such as the bivariate model of Blanchard and Quah (1989) the aggregate demand and aggregate supply shocks must be viewed as aggregates of a larger number of demand and supply shocks (see Faust and Leeper 1997). For example, in reality there may be labor supply shocks and productivity shocks rather than just one aggregate supply shock. As Blanchard and Quah (1989) point out, this fact may invalidate the economic interpretation of their shock estimates. For example, even if none of the underlying demand shocks affect real output in the long run, the estimated aggregate demand shock in their model will represent a mixture of *both* the underlying demand and supply disturbances. Blanchard and Quah (1989) provide a theorem clarifying when this problem does not occur, but the conditions underlying that theorem are quite restrictive.

Faust and Leeper (1997) demonstrate that, in general, one cannot extract aggregate demand shocks in Blanchard and Quah's 1989 bivariate model that only involve the underlying demand shocks, nor can one extract aggregate supply shocks that only involve the underlying supply shocks, because each of these shocks involves different dynamic responses. The source of the problem is that the DGP has more shocks than the estimated model, and that each shock triggers different dynamic responses. One potential solution is to estimate larger VAR models that allow for more shocks (see Faust and Leeper 1997; Erceg, Guerrieri, and Gust 2005). Indeed, many applied researchers have augmented the Blanchard and Quah (1989) model to include additional variables (see, e.g., Galí 1999). There is evidence, however, that the results may be sensitive to the choice of the additional variables. This means that it is difficult to draw general lessons from estimates of models based on long-run restrictions. Note, however, that omitted variables also distort impulse responses when short-run restrictions are used. In fact, much the same problem arises also in DSGE models. For example, an aggregate technology shock in an RBC model is merely a convenient fiction that obscures the fact that there are many potential sources of variation in aggregate technology with potentially different effects. The point of raising this issue in the context of structural VAR models with long-run restrictions is that it is important to be aware of the fact that this problem cannot be circumvented by using long-run restrictions.

10.5.3 Lack of Robustness at Lower Data Frequencies

It may seem that models based on long-run restrictions would apply equally regardless of the frequency of the data, making this approach particularly attractive when dealing with, say, quarterly or annual data, for which conventional short-run identifying assumptions are more difficult to justify. Faust and Leeper (1997) caution against this interpretation on the grounds that time aggregation will tend to invalidate the assumption of orthogonal structural shocks, even if that assumption applies at higher frequencies.

10.5.4 Nonuniqueness Problems without Additional Sign Restrictions

Estimates of the impulse responses in VAR models identified by short-run or long-run restrictions are identified only up to their sign. This fact matters both for the construction of impulse response point estimates and for the construction of simulated confidence intervals (see Chapter 12). In solving for the unknown elements of B_0^{-1}, it is typically implicitly assumed that the k^{th} shock has a positive effect on the k^{th} variable. There are situations in which such an assumption is natural. For example, in typical semistructural VAR models of monetary policy we would expect a contractionary monetary policy shock to be associated with a higher interest rate. Likewise, fully structural macroeconomic VAR models based on short-run restrictions can usually be written such that the k^{th} shock has a positive effect on the k^{th} variable (see Taylor 2004).

In other situations, the normalization can be less clear. This is especially true for models identified based on long-run restrictions. Consider, for example, the model of Blanchard and Quah (1989). Recall that $B_0^{-1} = \mathbf{\Gamma}(1)\mathbf{\Upsilon}$, where

$$\mathbf{\Upsilon} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} = \text{chol}(\mathbf{\Gamma}(1)^{-1}\Sigma_u\mathbf{\Gamma}(1)'^{-1})$$

and $*$ denotes an unrestricted element. The solution $\mathbf{\Upsilon}$ is unique due to the uniqueness of the Cholesky decomposition. However, for any positive definite matrix Ω, one may reverse the signs of all elements in any column of $\text{chol}(\Omega)$ and still preserve $\Omega = \text{chol}(\Omega)\text{chol}(\Omega)'$. Thus, as Taylor (2004) notes, the solution for B_0^{-1} is identified only up to a transformation. Equivalent solutions may be obtained by multiplying any of the columns of B_0^{-1} by -1, resulting in 2^K possible solutions, all of which satisfy $B_0^{-1}B_0^{-1\prime} = \Sigma_u$ (see also Lütkepohl

2013a). The practical effect of flipping the sign is to flip the structural impulse response functions in question about the horizontal axis. An obvious implication is that users of VAR models based on long-run restrictions need to make explicit which additional sign restrictions they are imposing for identification.

To obtain a unique solution, in practice, we need additional information from economic theory about the sign of the short-run or the long-run response. In some cases, this information is obvious. For example, Taylor (2004) discusses a bivariate VAR model of U.S. real GDP and its implicit price deflator, in which aggregate supply shocks have permanent effects on real GDP, but aggregate demand shocks do not. Of the $2^2 = 4$ possible solutions for B_0^{-1}, three can be immediately ruled out because we know that aggregate demand shocks move prices and quantities in the same direction on impact, while aggregate supply shocks move them in opposite directions. In other cases, identification may be less straightforward. For example, economic theory may not be informative about the sign of the impact response in question, in which case it is unclear how to proceed.

A case in point is the debate about the sign of the impact response of real output to a productivity shock, with some economists suggesting an initial decline and others an initial increase. Long-run restrictions were considered appealing in this context, precisely because of the perception that they leave short-run responses unrestricted. Without an additional normalization, however, models based on long-run restrictions cannot answer this question, and choosing the normalization based on the sign of the short-run response of real GDP simply amounts to assuming the answer. In this context, a more appealing strategy therefore is to normalize the sign of the response function based on the sign of the long-run response of real output. A similar concern also arises in the debate on the liquidity effect (see Section 8.5.3).

This problem is also important when conducting inference on models with long-run restrictions. In this case, an obvious concern is that, without an explicit sign assignment at each iteration, bootstrap replications of the model solution may correspond to different sign normalizations, invalidating inference (see Chapter 12). Moreover, Lütkepohl (2013a) observes that if an explicit sign assignment is carried out on an impact response coefficient that is close to zero, additional problems arise because the estimated sign need not coincide with the actual sign. For example, if one of the diagonal elements of B_0^{-1} is zero in population, and we normalize its estimate to be positive, then we will make the other elements flip their sign with positive probability regardless of their true value. This fact will inflate the width of bootstrap confidence intervals. Lütkepohl (2013a) illustrates this point in the context of the Blanchard and Quah (1989) model. In this model, the impact response of real output to an aggregate supply shock is close to zero. Normalizing the response on this coefficient dramatically inflates the width of bootstrap confidence intervals and changes the statistical significance of the impulse response estimates compared with normalizing on the unemployment response.

10.5.5 Sensitivity to Data Transformations

It has been observed that the conclusion from Blanchard-Quah type VAR models are sensitive to whether the second variable (e.g., hours worked) is entered in levels or differences. For example, specifying a VAR model with both hours worked and labor productivity in differences, Galí (1999) finds that hours worked initially drop after a positive technology shock, a finding that lends support to models with embedded frictions. On the other hand, Christiano, Eichenbaum, and Vigfusson (2004) provide support for the predictions of standard real business cycle (RBC) models, with hours worked rising immediately after a positive productivity shock, using the same long-run identification scheme, but allowing hours worked to enter the model in levels (see Section 11.5).

Gospodinov, Maynard, and Pesavento (2011) clarify the source of the extensive debate on the effect of technology shocks on unemployment/hours worked that ensued from these conflicting empirical results. They find that the contrasting conclusions from specifying the second VAR variable in levels as opposed to differences can be explained by small, but important, low-frequency co-movement between hours worked and labor productivity, which is allowed for in the levels specification but is implicitly set to zero in the differenced specification. Their theoretical analysis shows that, even when the root of hours is very close to one and the low-frequency co-movement is quite small, assuming away or explicitly removing the low-frequency component can have important implications for the long-run identifying restrictions, giving rise to biases large enough to account for the empirical difference between the two specifications. We defer a more formal discussion of this problem to Chapter 11. For a closely related analysis, see also Canova, López-Salido, and Michelacci (2010).

Which specification is right is ultimately an economic question and continues to be debated. For example, Fernald (2007) makes the case that the observed low-frequency correlation in the data is spurious and arises from breaks in both productivity and hours worked in the early 1970s and mid-1990s. Francis and Ramey (2009) instead attribute the observed correlation to common low-frequency trends in demographics and in public employment that are beyond the scope of the economic model. This view implies that the low-frequency component ought to be removed prior to the analysis, which leads to results that support the earlier findings of Galí (1999). On the other hand, if there is a true low-frequency correlation in the population model, as maintained by Christiano, Eichenbaum, and Vigfusson (2004), then any procedure that removes the low-frequency correlation between hours and productivity, whether by differencing, HP-filtering, or removing a deterministic time trend with breaks, will result in substantial bias in the estimates.