Measurement With Minimal Theory: Estimating The Dynamic Effects of Shocks

Karel Mertens, Cornell University

Contents

References

Bernanke, B. S. and Mihov, I. (1998), 'Measuring monetary policy', *Quarterly Journal of Economics* **113**(3), 869–902.

Blanchard, O. and Perotti, R. (2002), 'An empirical characterization of the dynamic effects of changes in government spending and taxes on output', *Quarterly Journal of Economics* **117**(4), 1329–1368.

Blanchard, O. and Quah, D. (1989), 'The dynamic effects of aggregate demand and supply disturbances', *American Economic Review* **79**(4).

Canova, F. (2007), *Methods for Applied Macroeconomic Research*, Princeton University Press.

Chari, V., Kehoe, P. and McGrattan, E. (2005), 'A critique of structural VARs using real business cycle theory'. Federal Reserve Bank of Minneapolis, Working Paper 631.

Chari, V., Kehoe, P. and McGrattan, E. (2007), 'Are structural VARs with long-run restrictions useful in developing business cycle theory?'. Federal Reserve Bank of Minneapolis, Staff Report 364.

Christiano, L., Eichenbaum, M. and Vigfusson, R. (2006), 'Assessing structural VARs'. in: Acemoglu, D., Rogoff, K. and Woodford, M. (Eds.), NBER Macroeconomics Annual 2006.

Christiano, L. J., Eichenbaum, M. and Evans, C. L. (1998), 'Monetary policy shocks: What have we learned and to what end?'. Handbook of Macroeconomics, ed. John B. Taylor and Michael Woodford eds. Vol 1A, 65-148.

Fernandez-Villaverde, J., Rubio-Ramirez, J. F., Sargent, T. J. and Watson, M. (2008), 'A,B,C's (and D's)'s for understanding VARs', *manuscript* .

Friedman, M. and Schwartz, A. J. (1963), *Monetary History of the United States, 1867-1960*, Princeton University Press.

Gali, J. (1999), 'Technology, employment, and the business cycle: Do technology shocks explain aggregate fluctuations?', *American Economic Review* **89**(1), 249–71.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Kascha, C. and Mertens, K. (2008), 'Business cycle analysis and VARMA models'. Norges Bank working paper.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer.

Perotti, R. (2007), 'Estimating the effects of fiscal policy in oecd countries', *manuscript* .

1 Vector Autoregressions

In this chapter we will focus on Vector Autoregressions (VARs), a class of statistical models that capture the statistical properties of many macroeconomic time series without relying explicitly on a particular economic theory. VARs are reduced form models, and it is therefore impossible to interpret the dynamics estimated using a VAR without at least some theoretical restrictions. Under certain conditions, it is possible to establish a link between VAR models and DSGE models, such as the RBC and monetary models we have seen in previous chapters. Conditional on selecting some meaningful identification restrictions, one can conduct economically interesting analysis employing (structural) VARs. These restrictions are potentially consistent with a wide variety of theoretical models and can be used to estimate structural impulse responses to productivity, monetary or other shocks or to make a decomposition of the contribution of the different shocks to aggregate volatility. Good sources for background reading are Hamilton (1994) and Lütkepohl (2005).

1.1 Some Preliminaries on Multivariate Time Series Econometrics

A vector autoregressive model of order p or *VAR(p) model* is

$$z_t = A_0 + A_1 z_{t-1} + ... + A_p z_{t-p} + u_t \ , \ t = 0, \pm 1, \pm 2, ... \tag{1}$$

where $z_t = \left[z_t^1, ..., z_t^n\right]'$ is a $n \times 1$ random vector, A_0 is an $n \times 1$ matrix of constants and $A_i, i = 1, ..., p$ are $n \times n$ fixed coefficient matrices and u_t is vector white noise satisfying $E\left[u_t\right] = 0$, $E\left[u_t u_t'\right] = \Sigma$, $E\left[u_t u_{t-j}'\right] = 0$ for $j \neq 0$ where Σ is a non-singular $n \times n$ matrix.[1]

A VAR(p) is *stable* if its reverse characteristic polynomial has no roots in and on the complex unit circle, i.e. if

$$\det\left(I_n - A_1\lambda - ... - A_p\lambda^p\right) \neq 0 \text{ for } \mid \lambda \mid \leq 1$$

A stable VAR(p) process $z_t, t = 0, \pm 1, \pm 2, ...$ is *stationary*, i.e. its first and second moments are time invariant.

[1] We will assume that u_t is *standard white noise*. u_t is standard white noise if the u_t are continuous random vectors satisfying $E\left[u_t\right] = 0$, $E\left[u_t u_t'\right] = \Sigma$, $E\left[u_t u_{t-j}'\right] = 0$ for $j \neq 0$ and, for some finite constant c:

$$E \mid u_t^i u_t^j u_t^k u_t^m \mid \leq c, \text{ for } i, j, k, m = 1, ..., n \text{ and all } t$$

A *lag operator* or *backshift operator* L is defined such that $Lz_t = z_{t-1}$. Using this operator, we can write (1) as

$$z_t = A_0 + A_1 L z_t + ... + A_p L^p z_t + u_t$$

or

$$A(L) z_t = A_0 + u_t \tag{2}$$

where $A(L) = I_n - A_1 L - ... - A_p L^p$. Define the inverse of $A(L)$ as $\Phi(L) = \sum_{i=0}^{\infty} \Phi_i L^i$ such that

$$\Phi(L) A(L) = I_n$$

Premultiplying (2) by $\Phi(L)$ gives the infinite order Moving Average or *MA(∞) representation*

$$z_t = \Phi(L) A_0 + \Phi(L) u_t = \left(\sum_{i=0}^{\infty} \Phi_i \right) A_0 + \sum_{i=0}^{\infty} \Phi_i u_{t-i}$$

Wold Decomposition Theorem Every $n \times 1$ stationary process z_t can be written as the sum of two uncorrelated processes $z_t = d_t + x_t$ where d_t is a deterministic process that can be forecast perfectly from its own past and x_t is a process with MA representation

$$x_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i}$$

where $\Phi_0 = I_n$ and u_t is a white noise process. We will usually consider systems in which the only deterministic component is the mean term, i.e. $d_t = d$ is a vector of constants. Under the condition that $\Phi(L)$ is invertible, $A(L) = \Phi(L)^{-1}$ and we can write z_t as a VAR of possibly infinite order:

$$\begin{aligned} x_t &= \sum_{i=1}^{\infty} A_i x_{t-i} + u_t \\ z_t &= A_0 + \sum_{i=1}^{\infty} A_i z_{t-i} + u_t \end{aligned}$$

where $A_0 = A(L)d$. In other words, under very general conditions, every stationary process can be written as a VAR process of infinite order. When the A_i's converge to zero sufficiently rapidly, every stationary process can be approximated well by a finite order VAR process.

1.2 Estimation of VARs

Estimating a VAR of order p is extremely straightforward and can be done by *multivariate least squares* or, after assuming a statistical distribution for the residuals u_t, by standard *maximum likelihood* procedures.

1.2.1 Multivariate Least Squares Estimator

The following discussion is based on Lütkepohl (2005). Suppose we have a sample $z_{t-p+1}, ..., z_T$ of stationary time series available, where $T + p$ is the total sample size. Let z be a $n \times T$ matrix defined as

$$z = \begin{bmatrix} z_1 & ... & z_T \end{bmatrix}$$

Let Z_t denote an $(np + 1) \times 1$ vector containing a constant term and p lags of each of the elements of z_t

$$Z_t = \begin{bmatrix} 1 \\ z_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ z_{t-p} \end{bmatrix}$$

Let Z be an $(np + 1) \times T$ vector collecting the T available observations of Z_t:

$$Z = \begin{bmatrix} Z_1 & ... & Z_T \end{bmatrix}$$

Similarly, we can collect the residual terms in the $n \times T$ matrix:

$$u = \begin{bmatrix} u_1 & ... & u_T \end{bmatrix}$$

Let A be an $n \times (np + 1)$ matrix defined as

$$A = \begin{bmatrix} A_0 & A_1 & ... & A_p \end{bmatrix}$$

Now define the $nT \times 1$ vector $\mathbf{z} = vec(z)$, the $(n^2T + n) \times 1$ vector $\beta = vec(A)$ and the $nT \times 1$ vector $\mathbf{u} = vec(u)$ where vec is the column stacking operator. We can now write (1) compactly as

$$z = AZ + u \tag{3}$$

or equivalently

$$\mathbf{z} = (Z' \otimes I_n)\beta + \mathbf{u} \tag{4}$$

where $\otimes$ denotes the Kronecker product. The covariance matrix of $\mathbf{u}$ is

$$\Sigma_u = I_T \otimes \Sigma$$

The Generalized Least Squares (GLS) estimator minimizes[2]

$$\begin{aligned}
\mathbf{u}'\Sigma_u^{-1}\mathbf{u} &= (\mathbf{z} - (Z' \otimes I_n)\beta)'\Sigma_u^{-1}(\mathbf{z} - (Z' \otimes I_n)\beta) \\
&= \mathbf{z}'\Sigma_u^{-1}\mathbf{z} + \beta'(Z \otimes I_n)\Sigma_u^{-1}(Z' \otimes I_n)\beta - 2\beta'(Z \otimes I_n)\Sigma_u^{-1}\mathbf{z} \\
&= \mathbf{z}'(I_T \otimes \Sigma^{-1})\mathbf{z} + \beta'(ZZ' \otimes \Sigma^{-1})\beta - 2\beta'(Z \otimes \Sigma^{-1})\mathbf{z}
\end{aligned}$$

The first order condition is

$$2(ZZ' \otimes \Sigma^{-1})\beta - 2(Z \otimes \Sigma^{-1})\mathbf{z} = 0$$

The GLS estimator $\hat{\beta}$ is therefore given by[3]

$$\begin{aligned}
\hat{\beta} &= ((ZZ')^{-1} \otimes \Sigma)(Z \otimes \Sigma^{-1})\mathbf{z} \\
&= ((ZZ')^{-1}Z \otimes I_n)\mathbf{z}
\end{aligned}$$

It can be shown (see Lütkepohl (2005) p.73-75) that under our assumptions the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta)$ is normal with covariance matrix $\Gamma^{-1} \otimes \Sigma$, i.e.

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \otimes \Sigma)$$

where $\Gamma = plim\frac{ZZ'}{T}$. In order to assess the asymptotic dispersion of the LS estimator, we still need to know Γ and Σ. Consistent estimators are given by

$$\begin{aligned}
\hat{\Gamma} &= \frac{ZZ'}{T} \\
\hat{\Sigma} &= \frac{1}{T - np - 1} z(I_T - Z'(ZZ')^{-1}Z)z'
\end{aligned}$$

[2]**Matrix Algebra:** Note that for general matrices X_1, X_2, X_3, X_4 of suitable dimensions,

- $(X_1 \otimes X_2)(X_3 \otimes X_4) = X_1X_3 \otimes X_2X_4$
- $(X_1 \otimes X_2)' = X_1' \otimes X_2'$
- $X_1 \otimes (X_2 + X_3) = X_1X_2 \otimes X_1X_3$
- if X_1, X_2 invertible: $(X_1 \otimes X_2)^{-1} = X_1^{-1} \otimes X_2^{-1}$

[3]The Hessian $2(ZZ' \otimes \Sigma^{-1})$ is positive definite such that $\hat{\beta}$ is a minimum.

The ratio $\frac{\hat{\beta}_i - \beta_i}{\hat{s}_i}$ has an asymptotic standard normal distribution. Here $\beta_i(\hat{\beta}_i)$ is the i-th element of $\beta(\hat{\beta})$ and $\hat{s}_i$ is the square root of the i-th diagonal element of

$$(ZZ')^{-1} \otimes \hat{\Sigma}$$

In other words, we can use the standard t-ratios in setting up confidence intervals and tests for individual coefficients. In macroeconomics however, it is rare to report results on the estimated VAR coefficients, because 1) they usually have little economic meaning, 2) they are poorly estimated in finite samples with large standard errors and 3) there is usually a large number $(n(n+1))$ of them. It is much more common to report functions of the VAR coefficients which may be more informative, have some economic interpretation and are hopefully better estimated. Examples are impulse response functions, forecast error variance decompositions and historical decompositions (see Canova (2007)). Discussion of these is deferred to the section on structural VARs.

1.2.2 Alternative Estimators

EXERCISE: Show that

1. The GLS estimator is identical to the multivariate OLS estimator obtained from minimizing $\mathbf{u}'\mathbf{u}$.

2. Multivariate GLS estimation is identical to applying the OLS estimator to each of the n equations in (1) separately.

3. The GLS estimator is identical to the maximum likelihood estimator assuming normal disturbances.

1.3 The ABC's and (D's) for Understanding VARs

This section discusses the link between VAR models and theoretical DSGE models. Recall that the general solution of a (linearized) DSGE model can be written as

$$s_{t+1} = Gs_t + Fe_{t+1} \tag{5}$$

$$z_t = Hs_t \tag{6}$$

where s_t is a $m \times 1$ vector of state variables, e_{t+1} is an $l \times 1$ vector of exogenous disturbances, z_t is an $n \times 1$ vector of variables of interest, G is $m \times m$, F is $m \times l$ and H is $n \times m$. m is

the number of state variables, l is the number of exogenous shocks and n is the number of variables of interest. It is important to remember that H and G are nontrivial functions of the model's structural parameters, such as β, δ, α, ... The shocks e_t are Gaussian vector white noise satisfying $E\left[e_t\right] = 0$, $E\left[e_t e_t'\right] = I$, $E\left[e_t e_{t-j}'\right] = 0$ for $j \neq 0$.
Substituting (5) in (6), the *state space representation* is

$$s_{t+1} = Gs_t + Fe_{t+1} \tag{7}$$

$$z_{t+1} = Cs_t + De_{t+1} \tag{8}$$

where $C = HG$ is $n \times m$ and $D = HF$ is $n \times l$. If s_t contained only variables that are observable to the econometrician, it would be straightforward to estimate any of the equations in (7) and (8) using OLS. In most macroeconomic applications however, s_t contains variables that are imperfectly or not observed to the econometrician, e.g. the capital stock, previous period productivity, ..., whereas z_t contains observable variables such as output, consumption, hours, investment,...
Under certain regularity conditions, it is possible however to recover a time series representation in terms of observables only. A first condition is

Stochastic Nonsingularity Condition: D is an $n \times n$ invertible matrix.

Note that stochastic nonsingularity implies $l = n$, i.e. the number of exogenous shocks equals the number of variables of interest n. When D is nonsingular, we can rewrite (8) as

$$e_{t+1} = D^{-1}(z_{t+1} - Cs_t)$$

Substituting this expression into (7), we have

$$s_{t+1} = \left(G - FD^{-1}C\right) s_t + FD^{-1} z_{t+1}$$
$$\Leftrightarrow \left[I - \left(G - FD^{-1}C\right) L\right] s_{t+1} = FD^{-1} z_{t+1}$$

where L is the lag operator.

Invertibility Condition: The eigenvalues of $G - FD^{-1}C$ are strictly less then one in modulus.

If the above invertibility condition holds, then we have

$$s_{t+1} = \left[I - \left(G - FD^{-1}C\right)L\right]^{-1} FD^{-1} z_{t+1} \tag{9}$$

We can shift (9) back one period and substitute into (8):

$$\begin{aligned} z_{t+1} &= C\left[I - \left(G - FD^{-1}C\right)L\right]^{-1} FD^{-1} z_t + De_{t+1} \\ z_{t+1} &= C\sum_{i=0}^{\infty}\left(G - FD^{-1}C\right)^i FD^{-1} z_{t-i} + De_{t+1} \end{aligned} \tag{10}$$

Equation (10) defines a VAR(∞) of observable variables as in (1) where

$$\begin{aligned} A_0 &= 0 \\ A_i &= C\left(G - FD^{-1}C\right)^i FD^{-1} \text{ for } i = 1, ..., \infty \\ u_t &= De_t \end{aligned}$$

Therefore, as long as the stochastic nonsingularity condition and the invertibility condition holds, there exists a VAR time series representation of a theoretical model with a solution given by (5)-(6). It is worth stressing several complications that arise in practice when using DSGE models to make the link between theory and VARs explicit:

1. The VAR representation of the model dynamics is usually of infinite order, such that with a finite sample of observations, the econometrician must chose a finite lag order when estimating (10). As noted above, as long as the A_i's converge to zero sufficiently rapidly, the model time series behavior can be approximated well by a finite order VAR process. However, as some recent research has recently stressed, it is important to check that this is indeed the case (see Chari, Kehoe and McGrattan (2005) among many others).

2. In practice, only a subset of equations of endogenous variables is used in VARs. Including too many variables increases the number of VAR coefficients to be estimated involving a loss in degrees of freedom; including too few may imply a loss of variables that are very informative for the dynamics of the time series.

3. The number of shocks in the model (l) and in the VAR (n) often differ, leading to problems of stochastic singularity. In practice, one has to add shocks to the model or assume some measurement error in the VAR specification.

4. The VAR representation is based on linearized versions of theoretical models. When a (log)linearization is a poor approximation to the theoretical model, so is a linear time series model. Nonlinear time series models are available, but are outside the scope of these lecture notes.

Finally, it is worth pointing out that there exist other, equivalent time series representations to theoretical models. Some alternative estimation algorithms are based on the state space representation, others on a VARMA representation (see Kascha and Mertens (2008)).

1.4 Structural VARs

We now know how to estimate the A_i coefficient matrices and Σ, the covariance matrix of the reduced form residuals of the VAR model. Henceforth the notation A_i, Σ refers to the estimates of these matrices. Note from the previous section that the relationship between the VAR residuals and the model's structural shocks is given by

$$\Sigma = D\Sigma_e D'$$

where $\Sigma_e = E[e_t e_t']$ is the covariance matrix of the structural economic shocks (productivity, monetary, etc) is taken to be a diagonal matrix, i.e. the fundamental structural shocks are uncorrelated. Note also that without loss of generality, we can normalize $\Sigma_e = I_n$, such that

$$\Sigma = DD' \tag{11}$$

This link between the reduced form residuals and the structural economic shocks allows for a structural VAR analysis.

Structural Impulse Responses If we can solve equation (11) for the individual elements of D, then it is possible to obtain empirical estimates of the *structural impulse responses* to the shocks e_t and trace out the dynamics of the endogenous variables in z_t. The estimated response of z_t in period $t+k$ to an impulse in the j-th element of e_t at date t is

$$z_{t+k} - E_{t-1}\left[z_{t+k}\right] \quad = \quad \Phi_k D e_t^j \ , \ k = 0, 1, ...$$

where $E_{t-1}\left[z_{t+k}\right] = d$, e_t^j is a $n \times 1$ vector with the j-th element of e_t as the only non-zero element and the dynamic multipliers Φ_k are the coefficients of the corresponding MA representation of the model. These can be obtained from the recursion:

$$\Phi_k = \sum_{i=1}^{k} \Phi_{k-i} A_i \ , \ k = 1, 2, ...$$

starting with $\Phi_0 = I_n$ and setting $A_i = 0$ for $i > p$. Sometimes, it is more instructive to look at the *cumulative impulse responses*

$$z_{t+k} - E_{t-1}[z_t] \quad = \quad \left(\sum_{i=0}^{k} \Phi_i \right) D e_t^j \ , \ k = 0, 1, ...$$

For instance, one might be interested in the long run impact of an innovation in the money growth rate on inflation, i.e. in

$$\begin{aligned} \lim_{k \to \infty} \left(\sum_{i=0}^{k} \Phi_i \right) D e_t^j &= \Phi(1) D e_t^j \\ &= (I_n - A_1 - ... - A_p)^{-1} D e_t^j \end{aligned}$$

where $\Phi(1)$, i.e. the MA polynomial evaluated at one, is the *total long run impact matrix.*

Identification The major problem that arises in estimating structural impulse responses is that in equation (11) D cannot have more unknown parameters than Σ. Since Σ is a symmetric matrix by construction (it is a covariance matrix), it can be summarized by $n(n+1)/2$ distinct values. However, D generally has n^2 free parameters: we cannot solve for the elements of D and the structural shocks are not identified unless the following identification conditions hold

- **Sign condition:** the diagonal elements of D are strictly positive. This is without loss of generality since it only leads to the normalization that a positive innovation in one of the elements of e_t represents a positive shock to the corresponding element of z_t.
- **Order condition:** there are at least $n(n-1)/2$ identifying restrictions on the elements of D. If the number of restrictions is exactly $n(n-1)/2$, the model is *just identified.* If the number of restrictions exceeds $n(n-1)/2$, the model is overidentified.
- **Rank condition:** The matrix derivative with respect to D of the equations defined by (11) is of full rank. This condition rules out that any column of D can be written as a linear combination of another column of D.

See Hamilton (1994) p.332 for a more formal analysis. A model is not identified if there is more than one D that yields the same Σ, or equivalently, if there is more that one set of structural parameters that make up the matrix D that leads to exactly the same value of the likelihood function. In that case, it is impossible to discriminate between the different models. Note that the sign, order and rank condition are only valid for "local" identification. There may be different economic shocks or different models that are consistent with the same VAR residuals. Sometimes "global" identification can be shown analytically, but this can only be done on a case by case basis. There are many types of assumptions one can make to satisfy the order condition and achieve identification, and we discuss two of the most popular ones: short run timing restrictions and restrictions on the long-run impact of structural shocks. In the rest of this chapter, we will only discuss models that are just identified and that satisfy the sign and rank conditions, unless mentioned otherwise.

Short-Run Timing restrictions Timing restrictions amount to restricting the contemporaneous response to a shock of some of the variables to zero. Consider for instance a VAR with output growth Δy_t , inflation π_t, nominal interest rates i_t and money growth Δm_t, such that $z_t = [\Delta y_t, \pi_t, i_t, \Delta m_t]'$. Suppose we have a class of models in which output growth reacts contemporaneously to only to its own shock, inflation responds contemporaneously to output and money shocks, interest rates respond contemporaneously only to money shocks and money growth responds to all shocks. In this case we may have:

$$\begin{aligned}
u_t &= De_t \\
&= \begin{bmatrix} d_{11} & 0 & 0 & 0 \\ d_{21} & d_{22} & 0 & d_{24} \\ 0 & 0 & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix} \begin{bmatrix} e_t^y \\ e_t^\pi \\ e_t^i \\ e_t^m \end{bmatrix}
\end{aligned}$$

or equivalently

$$\begin{aligned}
u_t^y &= d_{11}e_t^y \\
u_t^\pi &= d_{21}e_t^y + d_{22}e_t^\pi + d_{24}e_t^m \\
u_t^i &= d_{33}e_t^i + d_{34}e_t^m \\
u_t^m &= d_{41}e_t^y + d_{42}e_t^\pi + d_{43}e_t^i + d_{44}e_t^m
\end{aligned}$$

We have $(n(n-1)/2) = 6$ zero restrictions so the model is identified and we can solve for the elements of D from (11). Now consider the alternative set of timing assumptions such that

$$u_t = \begin{bmatrix} d_{11} & 0 & 0 & 0 \\ d_{21} & d_{22} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 \\ d_{41} & d_{42} & d_{43} & d_{44} \end{bmatrix} \begin{bmatrix} e_t^y \\ e_t^\pi \\ e_t^i \\ e_t^m \end{bmatrix}$$

such that D is (lower) triangular. The system is in this case said to be *recursive* and D can found in a simple manner from the Choleski decomposition of Σ.[4]

Short-run timing restrictions of this type are often hard to justify from a theoretical point of view. At least at high enough frequency, short run restrictions may be sometimes be plausible. For instance with quarterly data, there may be no contemporaneous response of output if capital and labor input adjustments take more than one quarter; there may be no contemporaneous response of the price level because prices are sticky for at least one quarter; the period t shock may not be in the period t information set of the agents that decide on the period's value of the variable; and so on. Fortunately, if one is interested in only one particular shock, it sometimes suffices to have much less restrictions than would be required for identification of all the shocks. Suppose for instance that we can partition $z_t = [Z_t^1, \bar{z}_t, Z_t^2]'$ where Z_t^1 and Z_t^2 are $n_1 \times 1$, resp. $n_2 \times 1$ vectors of variables and $\bar{z}_t$ is the period t observation of a variable that is associated with the shock of interest. Christiano, Eichenbaum and Evans (1998) show that the following zero restrictions suffice in order to identify the contemporaneous impact of the shock:

$$D = \begin{bmatrix} \underset{n_1\times n_1}{d_{11}} & \underset{n_1\times 1}{0} & \underset{n_1\times n_2}{0} \\ \underset{1\times n_1}{d_{21}} & \underset{1\times 1}{d_{22}} & \underset{1\times n_2}{0} \\ \underset{n_2\times n_1}{d_{31}} & \underset{n_2\times 1}{d_{32}} & \underset{n_2\times n_2}{d_{33}} \end{bmatrix}$$

i.e. D is block triangular. Clearly the first n_1 and the last n_2 equations are indistinguishable within each block. But to estimate the impulse responses to the shock associated with $\bar{z}_t$, we only need the corresponding column of D to be identified, not all the elements of D. Christiano et al. (1998) prove that

[4] If A is a general positive definite $n \times n$ matrix, then there exists a lower triangular matrix P with positive main diagonal such that $P^{-1}A(P')^{-1} = I_n$ or A=PP'. This decomposition is sometimes called a *Choleski decomposition*.

1. There is a nonempty family of D matrices of the above form, one of which of course is the lower triangular matrix with positive diagonal elements, that satisfy $\Sigma = DD'$
2. Each member of this family generates precisely the same impulse response to the shock associated with $\bar{z}_t$.
3. Adopting the normalization of picking the lower triangular D, the impulse response associated with $\bar{z}_t$ is invariant to the ordering of variables within Z_t^1 and Z_t^2.

In other words, doing a Choleski decomposition of Σ to obtain the impulse response to one particular shock involves much less timing restrictions.

Long-run Impact Restrictions The timing restrictions above, as well as many other types of restrictions made directly on the elements of D are often not very appealing. Blanchard and Quah (1989) show that imposing restrictions on D directly is in fact not necessary for identification of the structural shocks. These authors consider the accumulated effects of shocks to the system and focus on the *total long run impact matrix* of the structural from error $\Phi^e(1)$:

$$\Phi^e(1) = (I_n - A_1 - ... - A_p)^{-1} D = \Phi(1) D$$

Blanchard and Quah (1989) impose zero restrictions on this matrix instead and therefore assume that some shocks do not have any total long-run effects. Consider their example of $z_t = [\Delta y_t, un_t]'$ where Δy_t is output growth and un_t is the unemployment rate. Demand shocks only have transitory effects on Δy_t and the accumulated long-run effect of such a shock on Δy_t is zero. If we place the demand shocks second, than this implies that

$$\Phi^e(1) = \begin{bmatrix} \phi_{11} & 0 \\ \phi_{21} & \phi_{22} \end{bmatrix} = \Phi(1) \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}$$

Note that

$$\Phi^e(1)\Phi^e(1)' = \Phi(1)\Sigma\Phi(1)'$$

were the right hand side is a function of the VAR estimates only. We can obtain $\Phi^e(1)$ from the Choleski decomposition of $\Phi(1)\Sigma\Phi(1)'$ and obtain the short-run impact matrix from

$$D = \Phi(1)^{-1}\Phi^e(1) \tag{12}$$

Long-run restrictions of the type suggested by Blanchard and Quah (1989) are often much more appealing from a theoretical point of view, as they often are consistent with a much larger class of models. However, recent research by Chari, Kehoe and McGrattan (2007) and Christiano, Eichenbaum and Vigfusson (2006) suggest that long-run identified VARs may perform poorly in some circumstances since small sample problems and lag truncation problems lead to poor estimates of D.

Note how the above short and long run identification schemes lead to an extremely straightforward estimation strategy: we can first estimate the A_i 's and Σ as outlined above and subsequently uncover the model dynamics from (11) or (12). Had we imposed restrictions on the VAR coefficients, joint estimation of the A_i's, Σ and D would be required using an appropriate method that imposes the parameter constraints during the estimation step.

To assess the statistical significance of the estimated structural impulse response, we need standard errors. Because the impulse response coefficients are complicated nonlinear functions of the VAR coefficients and the covariance matrix of the shocks, it is generally not easy to find their distribution. We will discuss two methods of computing confidence intervals for the impulse response coefficients: the *delta method* which is based on the asymptotic distribution of the impulse responses; another is based on the *bootstrap method*, which is more useful in small samples.

Distribution of Impulse Responses: Delta Method Suppose $\hat{\beta}$ is an estimator of the vector β with $\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta)$. If $g(\beta) = [g_1(\beta), ..., g_m(\beta)]'$ is a vector-valued continuously differentiable function with $\frac{\partial g}{\partial \beta'} \neq 0$ at β, then

$$\sqrt{T}\left(g(\hat{\beta}) - g(\beta)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(\beta)}{\partial \beta'} \Sigma_\beta \frac{\partial g(\beta)'}{\partial \beta}\right)$$

Impulse responses are differentiable functions of the VAR parameters and of the covariance matrix. Therefore their asymptotic distribution can be obtained by applying the delta method. However, as pointed out by Canova (2007), the delta method has three problems: 1) they have poor properties in experimental design with realistic sample sizes (i.e. 100 to 200 quarterly observations); 2) the asymptotic coverage is very poor when near-unit-roots or near singularities are present; 3) the standard errors tend to be very large, just as the standard errors of the VAR coefficients. Therefore methods that exploit the small sample properties of the VAR coefficients, such as bootstrapping, are usually preferred in practice.

Distribution of Impulse Responses: Bootstrap Method Bootstrapping is a method for estimating the unknown small sample distribution of an estimator, such as impulse responses coefficients, that uses resampling with replacement from the original sample. One example is the *residual bootstrap* (see also Lütkepohl (2005) p. 709-712):
Suppose we have a sample of $T+p$ observations on a vector of time series z_t:

1. Estimate the VAR(p) using all available data. Obtain the estimated coefficient matrices $\hat{A}_i$ and the estimated reduced form residuals $\hat{u}_t$, $t = 1, ..., T$. Compute centered residuals $\hat{u}_t - \bar{u}$ where $\bar{u} = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t$ is just the average.

2. Obtain bootstrap residuals u_t^*, $t = 1, ..., T$ by randomly drawing with replacement from the centered residuals.

3. Construct artificial time series z_t^* by iterating on

$$z_t^* = \hat{A}_1 z_{t-1}^* + ... + \hat{A}_p z_{t-p}^* + u_t^* \ , \ t = 1, ..., T$$

   using the same p initial values for each generated series.

4. Reestimate the VAR using the artificial time series and compute the impulse responses of interest.

5. Repeat step 2 to 4 N times, where N is a large number.

Confidence intervals and/or standard errors can be computed based on the distribution obtained from the N bootstrap replications. Note that the above procedure can easily be applied to obtain confidence bands for other statistics of interest. Sometimes a related method based on Monte Carlo simulations is used to construct confidence intervals. In this case, the reduced form residuals in each replication are sampled from an assumed distribution (usually multivariate normal).

Historical Decomposition There are other ways of presenting economically meaningful results of structural VARs, one of which is a historical decomposition. To verify what role the j-th shock plays in explaining the dynamics of z_t, it is informative to compute the time series $z_t(j)$ by iterating on

$$z_t(j) = A_0 + A_1 z_{t-1}(j) + ... + A_p z_{t-p}(j) + D^j e_t^j \tag{13}$$

for $t = 1, ..., T$ given p initial observations $z_{-p}(j) = z_{-p}, ..., z_0(j) = z_0$. e_t^j is the j-th shock of the estimated e_t and D^j is the j-th column of D.

Forecast Error Variance Decomposition Another way to assess the relative contribution of different structural shocks to fluctuations in z_t is to do a forecast error variance decomposition. In the context of the MA representation

$$z_t = d + \sum_{i=0}^{\infty} \Phi_i^e e_{t-i}$$

where $\Phi_i^e = \Phi_i D$ and with the normalization $\Sigma_e = I_n$, the error of the optimal h-step forecast of z_t is

$$z_{t+h} - z_t(h) = \sum_{i=0}^{h-1} \Phi_i^e e_{t+h-i}$$

Focusing on a specific variable m in z_t, z_t^m, the forecast error is

$$\begin{aligned} z_{t+h}^m - z_t^m(h) &= \sum_{i=0}^{h-1} [\phi_i^{m1} e_{1,t+h-i} + ... + \phi_i^{mn} e_{n,t+h-i}] \\ &= \sum_{l=0}^{n} [\phi_0^{ml} e_{l,t+h} + ... + \phi_{h-1}^{ml} e_{l,t+1}] \end{aligned}$$

where ϕ_i^{ab} is the ab-th element of Φ_i^e and $e_{b,t}$ is the b-th element of e_t. The *mean squared error* (MSE) of the forecast $z_t^m(h)$ is

$$E[\left(z_{t+h}^m - z_t^m(h)\right)^2] = \sum_{l=0}^{n} [(\phi_0^{ml})^2 + ... + (\phi_{h-1}^{ml})^2]$$

since the elements of e_t have unit variance and are uncorrelated. Therefore

$$(\phi_0^{ml})^2 + ... + (\phi_{h-1}^{ml})^2 = \sum_{i=0}^{h-1} (\phi_i^{ml})^2$$

can be interpreted as the contribution of innovations to the l-th shock to the h-step forecast of the m-th variable. The proportion of the h-step forecast error in variable m that is due to shock l is thus

$$\frac{\sum_{i=0}^{h-1} (\phi_i^{ml})^2}{\sum_{l=0}^{n} [(\phi_0^{ml})^2 + ... + (\phi_{h-1}^{ml})^2]}$$

This kind of analysis is also called *innovation accounting*.

The Narrative Approach Before moving on, it is worth briefly mentioning an alternative strategy to identifying structural shocks called the *narrative approach*, pioneered by Friedman and Schwartz (1963). Friedman and Schwartz looked at a number of historical events in which changes in the money supply seemed to have been exogenous events, e.g. because of changes in reserve requirements that were not implemented because of changes in economic conditions. By looking at a variety of historical sources, it may be possible to collect a number of such exogenous events that can be used for structural analysis. The narrative approach has been applied to identify monetary policy shocks, government spending shocks and tax shocks among others.

2 VAR-Based Results on the Dynamic Response to Shocks

Unfortunately, convincing identification restrictions are hard to come by. As Hamilton (1994) notes, if there were compelling identification assumptions, macroeconomists would have settled their disputes a long time ago. This section presents some influential examples of empirical studies employing various identification schemes to uncover the responses to technology, monetary and fiscal shocks.

2.1 The Dynamic Effects of Technology Shocks

In an influential paper, Gali (1999) estimates the dynamic response of output and hours worked to an identified technology shock in order to assess the role of technology shocks in explaining business cycle fluctuations. His motivation is the following: standard RBC models are quite successful in matching many of the unconditional moments of US postwar data. As we have seen before, one dimension in which the simple RBC model does poorly is the prediction of a very high correlation between hours worked and labor productivity and real wages, whereas the correlation in the data is near-zero. Modifications that help to reconcile the RBC model with the data include adding additional shocks, such as government spending shocks or preference shocks. At that point, the RBC model yields predictions about *conditional* moments: i.e. conditional on a given source of fluctuations such as technology shocks. In order to falsify the model, it therefore becomes necessary to look at *conditional* moments in the data, which is exactly what structural VAR analysis is for. Gali's central objective is to verify whether the RBC model's predictions about what happens after a technology shock to output and hours are consistent with the data.

VAR specification Gali's empirical specification starts from the MA representation of possible infinite order:

$$z_t = \Phi(L)De_t$$

The vector of observables $z_t = [\Delta s_t, \Delta n_t]$ contains the first differences of the log of labor productivity $s_t = \log(S_t) = \log(Y_t/N_t)$ where Y_t is output and $n_t = \log(N_t)$ are log hours worked. The vector of structural shocks $e_t = [e_t^a, e_t^m]'$ contains a technology shock e_t^a and a non-technology shock e_t^m. Without loss of generality, we can impose the normalization that $E[e_t e_t'] = I_2$. The fact that the data are in first differences is based on an assumption that both labor productivity and hours are integrated of order 1, such that the first differences are stationary. Note that the assumption that labor productivity has a unit root in levels is certainly plausible, but the unit root in the level of hours is debatable (why?). Gali makes this assumption based on the outcome of unit root tests, but also employs an alternative empirical model with $z_t = [\Delta s_t, \bar{n}_t]$ where $\bar{n}_t$ denotes deviations of log hours from a fitted linear trend.

Identification The identification of the structural shocks in e_t is based on three assumptions:

1. Output is determined according to a homogenous of degree one, strictly concave, aggregate production function

$$Y_t = F(K_t, X_t N_t^e)$$

   where K_t and $X_t N_t^e$ denote the effective capital and labor-input services employed. The term "effective" refers to the possibility that there may be unobservable variations in the rate of utilization in both inputs. X_t is an exogenous technology parameter following a stochastic process with a unit root, i.e. some technology shocks have permanent effect on the level of X_t.

2. The capital-labor ratio measured in efficiency units, $\frac{K_t}{X_t N_t^e}$, follows a stationary stochastic process.

3. Effective labor input N_t^e is a homogenous of degree one function of hours N_t and effort U_t:

$$N_t^e = g(N_t, U_t)$$

   and effort per hour $\frac{U_t}{N_t}$ follows a stationary stochastic process.

The first assumption implies that there is stochastic growth in the economy as opposed to the RBC model with deterministic growth we have covered in the lecture notes. Apart from that difference, assumption 1 contains just the same restrictions on the production function that are required for balanced growth. It is even more general than the model from the RBC chapter, because it allows for fluctuations in the rate of utilization of factor inputs. Assumption 1 is therefore very general and holds in a large class of models that display growth. The second assumption is also hardly restrictive because it implies that the return on capital is stationary. As long as they are consistent with balanced growth, almost all business cycle models have this feature. Moreover, in the data asset returns are found to be stationary. In assumption 3, homogeneity of the function g is required of effective labor input is to be proportional to hours whenever effort per hour is constant. Stationarity of $\frac{U_t}{N_t}$ seems empirically plausible and is consistent with all models with variable effort. Note that of course, assumption 3 also nests all models without variable labor effort, such as the models from the lecture notes. In short, the three assumptions hold for a large class of models and are not very restrictive.

Combining assumptions 1 to 3, the measured labor productivity can be written as

$$S_t = \frac{Y_t}{N_t} = \frac{Y_t N_t^e}{N_t N_t^e} = X_t F\left(\frac{K_t}{X_t N_t^e}, 1\right) g\left(1, \frac{U_t}{N_t}\right)$$

and after taking logs

$$s_t = x_t + \zeta_t \tag{14}$$

where $\zeta_t = \log\left(F\left(\frac{K_t}{X_t N_t^e}, 1\right) g\left(1, \frac{U_t}{N_t}\right)\right)$ is stationary under the three assumptions. Equation (14) is what allows Gali to identify technology shocks. It implies that *only* permanent changes in the stochastic component of the technology parameter x_t can be the source of the unit root in labor productivity. In other words, only permanent technology shocks can have a permanent effect on the level of labor productivity, even though other shocks may have transitory effects. Note that non-technology shocks can have a permanent impact on hours and output. Therefore, Gali's identification is different from Blanchard and Quah (1989) where supply shocks where defined more generally as the only shocks that have a permanent effect on output. Also keep in mind that transitory technology shocks will be captured by the non-technology shock. Formally, the assumptions implies that the long run impact matrix of the structural shocks, $\Phi^e(1)$ is lower triangular:

$$\Phi^e(1) = \begin{bmatrix} \phi_{11} & 0 \\ \phi_{21} & \phi_{22} \end{bmatrix} \quad = \quad \Phi(1) \begin{bmatrix} d_{11} & d_{21} \\ d_{21} & d_{22} \end{bmatrix}$$

After obtaining estimates of $\Phi(1)$ and Σ using a VAR, it is possible to construct the estimated impulse responses as outlined above.

Results Figure 1 displays the response of output and hours to a one standard deviation permanent positive technology shock for the model with first differenced hours. The estimates are based on the sample 1948Q1-1994:4 and the bands are $+/-$ two standard error bands computed using a Monte Carlo simulation of 500 replications. The results are in dramatic contrast with the central predictions of the RBC model: the estimated impulse responses show a persistent *decline* of hours worked in response to a *positive* technology shock. As is clear from the historical decomposition in Figure 2, movements in hours and output attributed to technology shocks are negatively correlated and therefore cannot account for much of the postwar US business cycles. The reverse is true for non-technology shocks. Gali concludes that the evidence is at odds with RBC theory, and presents a new Keynesian alternative that is largely consistent with his findings. Needless to say that Gali's results have been scrutinized and criticized along all possible dimensions, see for instance Chari, Kehoe and McGrattan (2005, 2007).

Figure 1: Response to a one-standard-deviation Positive Technology Shock, with two-standard-error bands (First-Differenced Hours), Gali (1999)

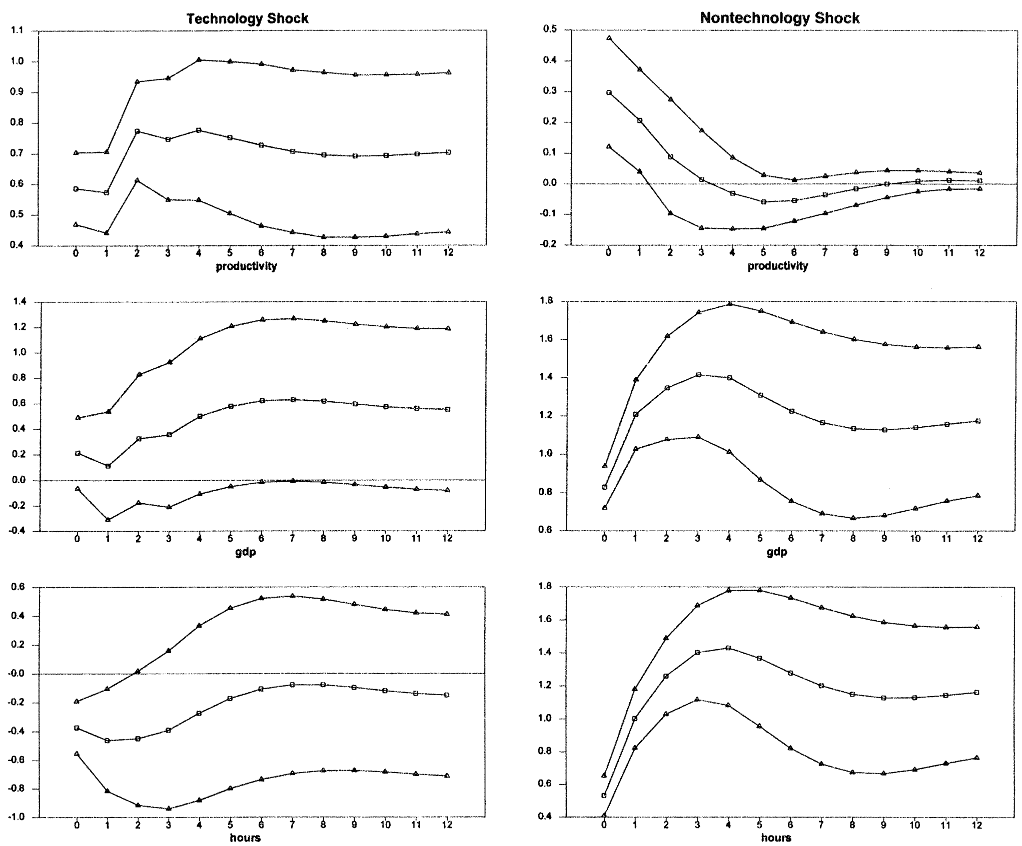
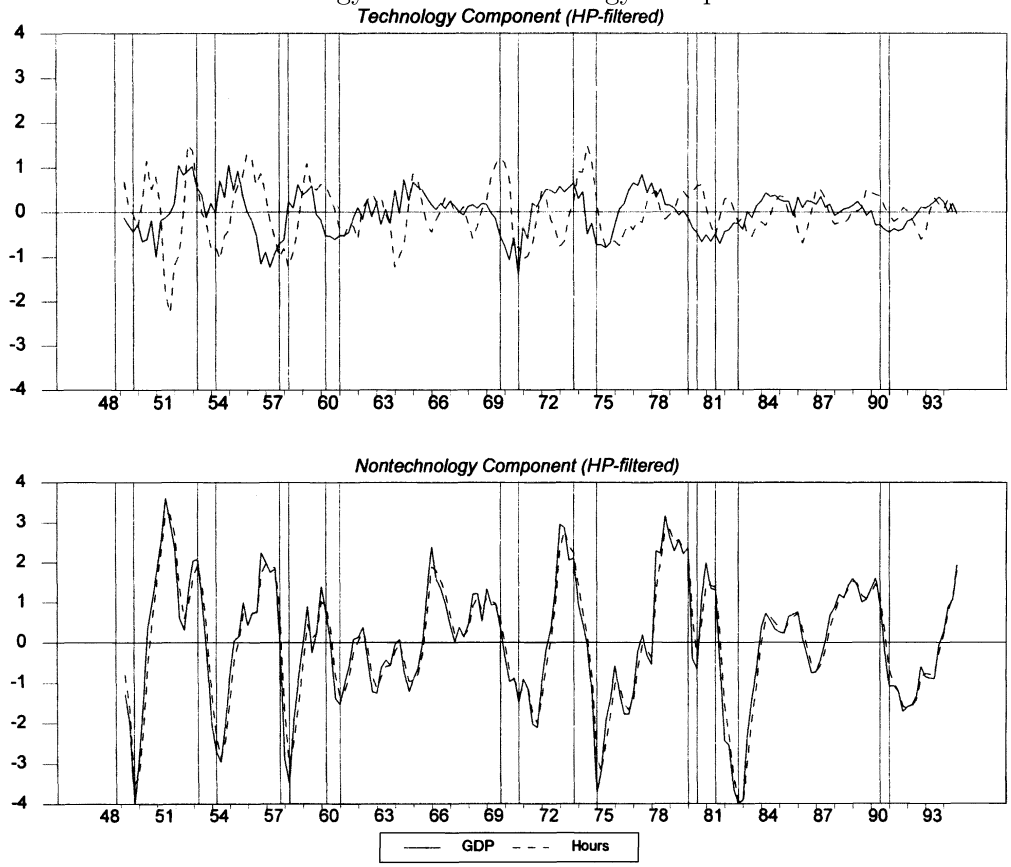


Figure 2: Estimated Technology and Non-Technology Components of US GDP and Hours.



2.2 The Dynamic Effects of Monetary Policy Shocks

Structural VARs have been very influential in the analysis of monetary policy and in the study of the monetary transmission mechanism. Of the numerous papers, this section will focus primarily on the contribution by Bernanke and Mihov (1998). A comprehensive overview of the empirical literature on monetary policy shocks is given by Christiano, Eichenbaum and Evans (1998).

Bernanke and Mihov provide a structural VAR methodology for measuring innovations in monetary policy and their effects on macroeconomic aggregates. One of the first complications that arises is how to measure monetary policy: the previous chapter on monetary models suggests to use money growth rates, which has been indeed the traditional approach. But there are many different monetary aggregates (M1,M2,...) one could use, and their growth rates have in many instances been very different. An even more fundamental problem is that growth rates of monetary aggregates depend on a variety of nonmonetary influences. Typically, the Federal Reserve has acted to keep nominal interest rates constant and has accommodated money demand shocks. Therefore, changes in money growth rates will often reflect changes in money demand as well as money supply. Moreover there have been dramatic changes in money velocity due to financial innovations that may lead us to confound changes in monetary policy with other shocks.

Bernanke and Mihov analyze several alternative measures for the stance of monetary policy that have been suggested in the more recent empirical literature: 1) the *federal funds rate*, 2) the *quantity of nonborrowed reserves* and 3) the *quantity of nonborrowed reserves growth that is orthogonal to total reserves growth*. The focus of their analysis is on exogenous innovations to these measures rather than the arguably more important systematic or endogenous component of monetary policy. Tracing the dynamic response to these innovations provides a means of observing the effects of policy changes under minimal identifying assumptions.

VAR specification The starting point is the following structural VAR model

$$\begin{aligned} z_t &= A_1(L)z_{t-1} + A_2(L)p_{t-1} + D_1 e_t + D_2 e_t^p \\ p_t &= A_3(L)z_{t-1} + A_4(L)p_{t-1} + D_3 e_t + D_4 e_t^p \end{aligned}$$

where z_t is a vector of macroeconomic variables, p_t is a vector of policy indicators. $A_1(L)$ to $A_4(L)$ are lag polynomials. e_t is a vector of nonmonetary structural shocks and e_t^p are shocks to the policy indicators, one of which is a money supply or monetary policy shock. The contemporaneous impact of the structural shocks are given by the coefficients in D_1 to D_4.

Identification The first identification assumption of Bernanke and Mihov is that D_2 is a matrix of zeros. This implies that none of the macroeconomic variables in z_t responds contemporaneously to shocks in either of the elements of e_t^p. If p_t is a scalar and contains only, say, the federal funds rate, than this is sufficient to estimate the impulse responses to a fed funds shock, which would automatically also be the monetary policy shock. This is because it constitutes a special case of the block-triangular D discussed in the section on timing restrictions. A simply Choleski decomposition of the estimated covariance matrix Σ in a VAR where the fed funds rate is ordered last is all that is needed.

However, Bernanke and Mihov invoke the following model of the market for bank reserves to elicit the assumptions needed to justify each of the three proposed policy indicators.

$$u_{TR} = -\alpha u_{FFR} + e^d \tag{15a}$$
$$u_{BR} = \beta(u_{FFR} - u_{DISC}) + e^b \tag{15b}$$
$$u_{NBR} = \phi^d e^d + \phi^b e^b + e^s \tag{15c}$$

Equation (15a) is the bank's total demand for reserves expressed in innovations form, i.e. an innovation in the demand for total bank reserves u_{TR} depends negatively on the innovation in the federal funds rate u_{FFR} and on a structural demand shock. Equation (15b) determines the portion of reserves that banks wish to borrow at the discount window u_{BR}, which depends negatively on the (innovation in the) discount rate and positively on the federal funds rate and on a structural shock e^b. The demand for nonborrowed reserves is of course given by $u_{TR} - u_{BR}$. Equation (15c) describes the supply of nonborrowed reserves by the Federal Reserve, which is allowed to depend on the demand shocks. The shock e^s is the monetary policy shock.

Letting $p_t = [TR_t, NBR_t, FFR_t]$ and making the simplifying assumption that $u_{DISC} = 0$ (see the paper for a justification), it is possible to solve the model as follows

$$\begin{bmatrix} u_{TR} \\ u_{NBR} \\ u_{FFR} \end{bmatrix} = \begin{bmatrix} -\frac{\alpha}{\alpha+\beta}(1-\phi^d)+1 & \frac{\alpha}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta}(1+\phi^b) \\ \phi^d & 1 & \phi^b \\ \frac{1}{\alpha+\beta}(1-\phi^d) & -\frac{1}{\alpha+\beta} & -\frac{1}{\alpha+\beta}(1+\phi^b) \end{bmatrix} \begin{bmatrix} e^d \\ e^s \\ e^b \end{bmatrix}$$

This system has seven unknown parameters, α, β, ϕ^d, ϕ^b and the three variances of the structural shocks, to be estimated from six covariances. The model is therefore underidentified by one restriction. Here we will discuss three indicators of monetary policy proposed in the literature and the corresponding identification restrictions they imply:

1) **The FFR model**: The federal funds rate reflects policy shocks, which requires $\phi^d = 1$ and $\phi^b = -1$ such that $u_{FFR} = -(\alpha+\beta)^{-1}e^s$.

2) **The NBR model**: Nonborrowed reserves reflect policy shocks, which requires $\phi^d = 0$ and $\phi^b = 0$ such that $u_{NBR} = e^s$.

3) **The NBR/TR model**: "Orthogonalized" nonborrowed reserves reflect policy shocks, which requires that $\alpha = 0$ and $\phi^b = 0$ such that $e^s = -\phi^d u_{TR} + u_{NBR}$

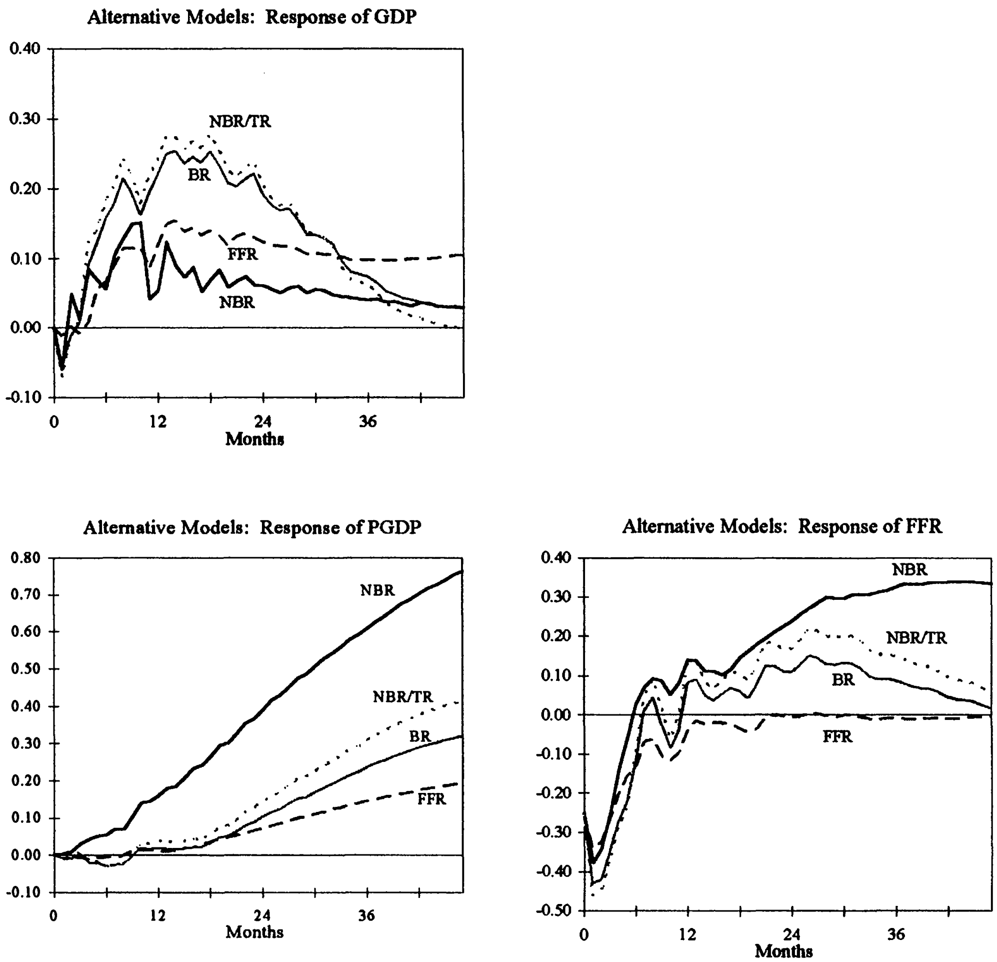
Note that each of the models imposes two additional identification restrictions, such that all of them lead to overidentification. This advantageous because the different models can be tested in the form of a test of the overidentifying restriction.

Results In the VAR, z_t includes real GDP, the GDP deflator, an index of spot commodity prices, p_t includes the federal funds rate, nonborrowed reserves and total reserves and the data are monthly and biweekly observations for the period 1965-1996. Figure 3 plots the estimated impulse responses of real GDP, the GDP deflator and with the federal funds rate to a shock with an impact response of -0.25% in the federal funds rate. An expansionary monetary shock increases output relatively rapidly and raises the price level more slowly but more persistently. Quantitatively, the results differ noticeable according the method of identifying policy shocks. Bernanke and Mihov also estimate a sizeable liquidity effect, here interpreted as the impact of a given increase in nonborrowed reserves on the interest rate.

Based on tests of the overidentifying restrictions, the FFR model is found to be (marginally) the best choice for measuring monetary policy shocks for the whole sample. One potential problem is that the Federal Reserve's operating procedure may have changed over time. Bernanke and Mihov test for this proposition and conclude that it has. The FFR model performs well for the 1965-1979 period and does exceptionally well for the Greenspan era, i.e. post 1988. However it does poorly for the Volcker era (1979-1982), for which the NBR model does well. The NBR model is strongly rejected in the other sample periods. The NBR/TR performs well over the various samples. Bernanke and Mihov conclude that the use of a single policy indicator for the entire period is not justified, but that the FFR model is probably a reasonable approximation for the greater proportion of the sample.

Overall, the results of Bernanke and Mihov and of many other studies using similar methodologies are not very favorable for the flexible price monetary models from the previous chapter. Christiano et al. (1998) give an overview of the results for a broader set of variables in the literature. The picture that emerges is that, after a contractionary monetary policy shock there is a sizeable reduction in various measures of economic activity and an increase in unemployment. There is generally a strong liquidity effect in the short run and the price level eventually declines but only after a very long period (about 1.5 years).

Figure 3: Response to a Monetary Policy Shock leading to a 0.25% reduction in the Federal Funds Rate, Bernanke and Mihov (1998)



2.3 The Dynamic Effects of Government Spending and Taxes

This section discusses the aggregate effects of fiscal policy shocks, such as changes in government spending and taxes. Perotti (2007) is a good source for a general overview of the empirical literature on fiscal policy shocks. We will focus here on a paper by Blanchard and Perotti (2002) that analyzes the effects of shocks to government spending and taxes using a SVAR approach in postwar US data. The authors achieve identification by making some timing assumptions and by exploiting institutional information about the tax and transfers systems and the timing of tax collections in the US.

VAR Specification The basic VAR specification is

$$z_t = B(L, q)z_{t-1} + u_t \tag{16}$$

where $z_t = [T_t, G_t, Y_t]'$ is a 3×1 vector with quarterly observations of T_t, the log of total tax revenues, G_t the log of total government spending and Y_t, the log of GDP, all in real, per capita terms. $u_t = [u_t^T, u_t^G, u_t^Y]'$ is the corresponding vector of reduced-form residuals. $B(L, q)$ is a four-quarter distributed lag polynomial that allows for the coefficients at each lag to depend on the particular quarter q that indexes the dependent variable. Blanchard and Perotti (2002) allow for quarter-dependence because of seasonal patterns in the response of some of the taxes to economic activity that are due to tax collection lags.

Identification Without loss of generality,

$$\begin{aligned} u_t^T &= a_1\epsilon_t^G + a_2 u_t^Y + \epsilon_t^T \\ u_t^G &= b_1\epsilon_t^T + b_2 u_t^Y + \epsilon_t^G \\ u_t^Y &= c_1 u_t^T + c_2 u_t^G + \epsilon_t^Y \end{aligned}$$

or equivalently

$$\begin{bmatrix} 1 & 0 & -a_2 \\ 0 & 1 & -b_2 \\ -c_1 & -c_2 & 1 \end{bmatrix} \begin{bmatrix} u_t^T \\ u_t^G \\ u_t^Y \end{bmatrix} = \begin{bmatrix} 1 & a_1 & 0 \\ b_1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_t^T \\ \epsilon_t^G \\ \epsilon_t^Y \end{bmatrix}$$

such that

$$D = \frac{1}{1 - b_2c_2 - c_1a_2} \begin{bmatrix} 1 - b_2c_2 + a_2c_2b_1 & a_1 - a_1b_2c_2 + a_2c_2 & a_2 \\ b_2c_1 + b_1 - b_1c_1a_2 & 1 + b_2c_1a_1 - c_1a_2 & b_2 \\ c_1 + c_2b_1 & c_1a_1 + c_2 & 1 \end{bmatrix}$$

where $\epsilon_t = [\epsilon_t^T, \epsilon_t^G, \epsilon_t^Y]'$ are the mutually uncorrelated structural shocks. Note for the sake of clarity that here the standard deviations of the structural shocks are not normalized to one, but instead the diagonal elements of D are. Identification requires finding values for a_1, a_2, b_1, b_2, c_1 and c_2. Blanchard and Perotti (2002) proceed in three steps:

1. They rely on institutional information about tax, transfer and spending programs to construct the parameters a_2 and b_2, which determine the response of T_t and G_t to unexpected movements in Y_t. In principle a_2 and b_2 capture two different effects of activity on spending and taxes: On the one hand there are the *automatic effects* of activity on T_t and G_t under the existing fiscal policy rules; On the other hand, there are the *discretionary adjustments* made to fiscal policy in response to unexpected events. One key timing assumption is that the use of quarterly data eliminates the second effect: it takes policymakers more than one quarter to learn about a GDP shock, decide upon an action and take it through the legislative process.

   Blanchard and Perotti (2002) could not find any systematic response of G_t to unexpected movements to GDP and set $b_2 = 0$. The choice of a_2 is based on the appropriate aggregation of elasticities of the various tax revenue components with (e.g. personal income taxes, social security taxes,...) with respect to their tax base, leading to $a_2 = 2.08$ for the sample 1947Q1-1997Q4.

2. The estimates a_2 and b_2 are used to construct the cyclically adjusted residuals $\tilde{u}_t^T = u_t^T - a_2u_t^Y$ and $\tilde{u}_t^G = u_t^G - b_2u_t^Y = u_t^G$, which are no longer correlated with ϵ_t^Y. They are therefore valid instruments to estimate c_1 and c_2 in a regression of u_t^Y on u_t^T and u_t^G.

3. The last two parameters a_1 and b_1 are the hardest to estimate convincingly and Blanchard and Perotti (2002) consider two alternative scenario's that differ on whether taxes are adjusted in response to decisions to change spending, or spending is adjusted in response to decisions to change taxes:

Case 1 $a_1 = 0, b_1 \neq 0$: Spending responds to increases in taxes.

Case 2 $b_1 = 0, a_1 \neq 0$: Taxes respond to increases in spending.

The values for a_1 resp. b_1 can be solved from the covariance matrix of the cyclically adjusted residuals.

Note that the identification assumptions lead to

$$D = \frac{1}{1 - c_1 a_2} \begin{bmatrix} 1 + a_2 c_2 b_1 & a_2 c_2 & a_2 \\ b_1 - b_1 c_1 a_2 & 1 - c_1 a_2 & 0 \\ c_1 + c_2 b_1 & c_2 & 1 \end{bmatrix} \qquad (\text{ Case } 1)$$

$$D = \frac{1}{1 - c_1 a_2} \begin{bmatrix} 1 & a_1 + a_2 c_2 & a_2 \\ 0 & 1 - c_1 a_2 & 0 \\ c_1 & c_1 a_1 + c_2 & 1 \end{bmatrix} \qquad (\text{ Case } 2)$$

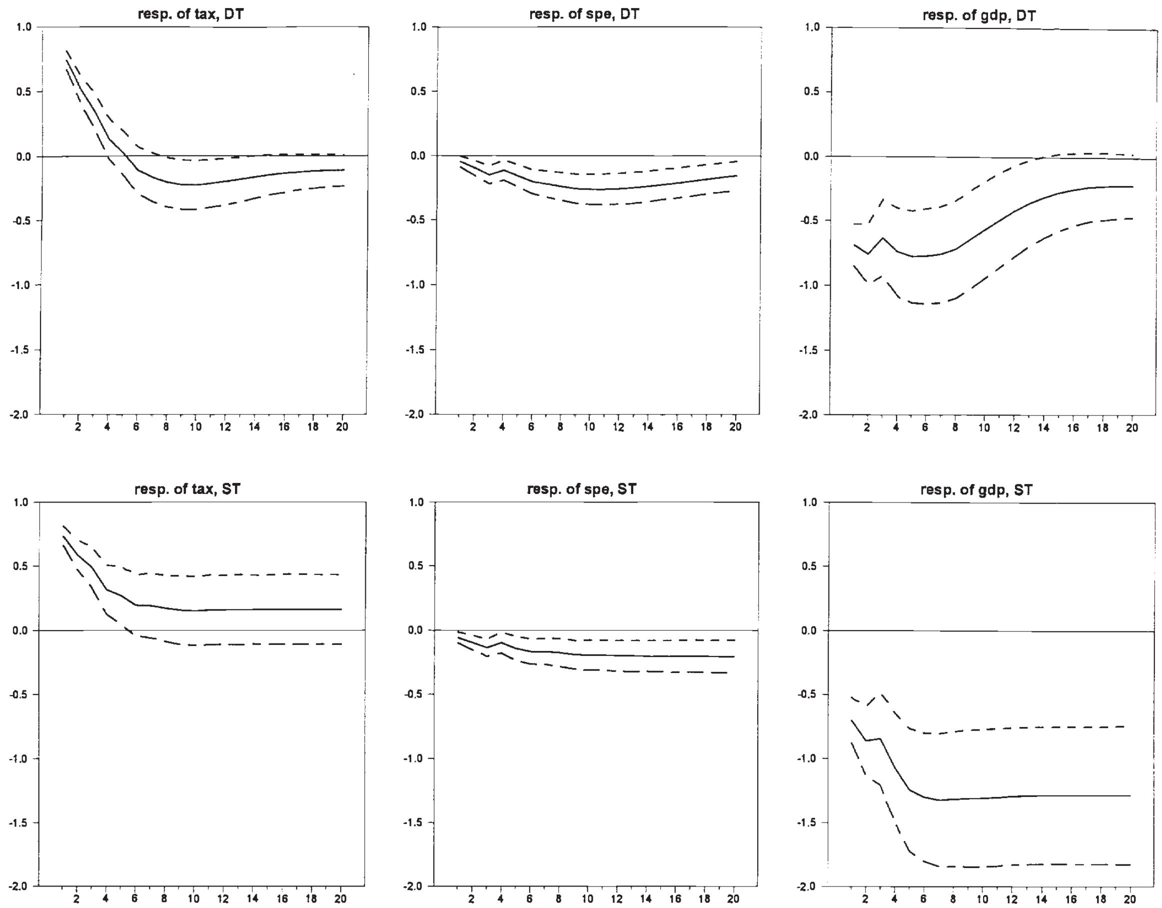
both of which satisfy the order condition for a given value of a_2.

There are a few other issues that complicate the analysis: The first is whether to assume a deterministic or stochastic trend.[5] Under the DT-specification, a deterministic trend is included by allowing for a linear and quadratic term in time in the VAR. The ST-specification instead assumes a stochastic trend and the VAR is conducted in first differences with an adjustment to account for changes in the underlying drift term. Second, a dummy variable is included for the exceptionally large tax cut in 1975:2. Finally, the quarter dependence in (16) implies that impulse responses are also quarter dependent. Blanchard and Perotti (2002) estimate the variance-covariance matrix using the quarter-dependent specification, but plot the dynamics that result from a VAR without quarter dependence to avoid having to present four different sets of results.

Results The results of Blanchard and Perotti (2002), displayed in Figures 4 and 5, show a consistently positive effect of government spending shocks on output, and a consistently negative effect of positive tax shocks on output. The graphs report point estimates of the impulse response coefficients as well as one standard error bands obtained from Monte Carlo simulations based on 500 replications. In their paper, the basic VAR is augmented to include the different components of GDP (one at a time). They find that a positive tax shock decreases consumption, investment and net exports, and that a positive spending shock increases consumption and net exports but leaves investment relatively unaffected. The positive consumption response to a spending increase is inconsistent with the neoclassical RBC model, but consistent with a standard Keynesian model; the implied negative investment response to a balanced budget expansion is consistent with the neoclassical RBC

[5] Formal statistical tests are inconclusive.

Figure 4: Response to a 1% Tax Increase (Case 1), with one standard deviation bands, Blanchard and Perotti (2002)



model, but hard to reconcile with a standard Keynesian model.

Figure 5: Response to a 1% Spending Increase (Case 2), with one standard deviation bands, Blanchard and Perotti (2002)