Structural VAR's*

1. Structural vs. Behavioral Models

The original meaning of a "structural" model in econometrics is explained in an article by Hurwicz (1962). A model is structural if it allows us to predict the effect of "interventions" — deliberate policy actions, or changes in the economy or in nature of known types. To make such a prediction, the model must tell us how the intervention corresponds to changes in some elements of the model (parameters, equations, observable or unobservable random variables), and it must be true that the changed model is an accurate characterization of the behavior being modeled after the intervention.

In the traditional simultaneous equations models that Hurwicz had in mind, the intervention was ordinarily taken to correspond to changing the parameters in an equation or block of equations in the model. The simplest conceptual example, corresponding to the monetary VAR literature, is where one block of equations describes policy behavior and another describes private sector behavior. The model is claimed to be structural because one set of policy equations can be replaced by another, while leaving the private sector equations unchanged, to obtain a prediction about the behavior of the economy with the new monetary policy.

However, there is no need for the intervention to correspond to changing an equation. In a model derived from a general equilibrium, for example, the natural parameters of the model (from utility functions, production functions, policy makers' objective functions) are likely to appear in many equations of the model. Such a model will claim to be structural relative to changes in at least some of these natural parameters — policy makers' objective functions, for example. One way to describe the Lucas critique of econometric policy advice is to say that he pointed out that parameters characterizing monetary policy behavior are likely to appear, via expectations, in many equations of the model, not just in the "policy equations". Thus an attempt to predict the effects of a policy change by changing only the policy equation, holding other equations in the model fixed, will fail, because the other equations will in fact change when the policy changes.

There is no sharp distinction among interventions that change equations, change parameters in equations, change disturbance terms in equations, or change the value of variables in the system. For example, in a monetary policy model there may be a reaction function describing monetary policy behavior, say

$$r_t = \alpha + X_t\beta + \varepsilon_t\,, \tag{1}$$

where X_t is a vector of explanatory variables. We might claim that the model is structural relative to changes in monetary policy, with these changes represented as changes in the monetary policy equation. Then we might ask what would be the distribution of the variables in the VAR system over the period t_0 to $t_0 + s$, conditional on policy setting r_t equal to some non-random fixed path $\bar{r}_t$ over this period. To do so, we could replace the policy equation (1) with the equation $r_t = \bar{r}_t$, or we could replace the time path of the disturbance to the policy equation with $\bar{\varepsilon}_t = \bar{r}_t - \alpha - X_t\beta$, or we could replace the fixed α and β in the equation with a sequence α_t, β_t satisfying $\bar{r}_t = \alpha_t + X_t\beta_t$ and set ε_t to zero. These would all deliver exactly the same implications for the behavior of the economy, because all retain the non-policy equations of the model unchanged, while fixing r_t at the $\bar{r}_t$ path.

Nowadays a model is often called "structural" when its parameters have behavioral interpretations, regardless of whether the old definition of the term applies, and on the other hand models that are in fact structural in the old sense are thought of as "reduced form" because they contain parameters or equations that do not have unique behavioral interpretations. Monetary policy VAR's, which single out a policy block and a non-policy block of equations, are certainly structural in the old sense (or at least claim to be), but because the separate equations in the non-policy block are often left uninterpreted, they are thought of as non-structural. Some real business cycle models, in contrast, are specified without explicit variables or equations representing monetary and fiscal policy, but are nonetheless calibrated to match some aspects of the behavior of macroeconomic data. There is no apparent interesting intervention with respect to which such models are structural in the original sense, but because all the parameters in the models have explicit behavioral interpretations, they are often referred to as structural.

My own preference is to reserve "structural" for its original meaning, and to use "behavioral" to characterize models with complete behavioral interpretations.

2. Structural VAR's and Simultaneous Equations Models (SEM's)

Both these classes of models can be thought of as versions of the general linear stochastic difference equation model

$$\underset{n\times n}{\Gamma}(L)y_t = c + \varepsilon_t\,, \tag{2}$$

where Γ is a matrix-valued polynomial in positive powers of the lag operator L and Γ_0 is full rank. The usual structural VAR framework specializes this setup by requiring that the elements of the ε_t vector be independent (in the Gaussian case that $\Sigma = \mathrm{Var}(\varepsilon_t)$ be diagonal). In most of the structural VAR literature it is assumed also that ε_t spans the space of the $y(t)$ innovation vector, i.e. that if we multiply the system through by Γ_0^{-1} to arrive at

$$\Gamma_0^{-1}\Gamma(L)y_t = B(L)y_t = \Gamma_0^{-1}c + \Gamma_0^{-1}\varepsilon_t = \gamma + \nu_t\,, \tag{3}$$

the result is the autoregressive representation of y, with $\Gamma_0^{-1}\varepsilon_t$ the innovation in y_t.

The usual SEM framework has two standard forms. In one, the system satisfies a Granger causal ordering, i.e. it can be written as

$$\begin{bmatrix} \Gamma_{11}(L) & \Gamma_{12}(L) \\ 0 & \Gamma_{22}(L) \end{bmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{xt} \\ \varepsilon_{zt} \end{bmatrix}, \tag{4}$$

with Σ block diagonal conformably with the x, z partition of the y vector. In this case the z's are called **exogenous**, or strictly exogenous. And in this case there is no requirement that ε_{xt} be an innovation. What is required is only that the full sample's $\{\varepsilon_{xt}\}$ vector be unrelated to (uncorrelated with, or independent of, depending on the context) the full sample's $\{z_t\}$ vector.

In the other standard form, only Γ_{110} is block triangular, Σ is block diagonal as before, and the ε_t vector is assumed to be the innovation vector. In this case $\{z_s \mid s \leq t\}$ are **predetermined**.

In both cases traditional treatments of the theory usually do not make explicit the z block of equations, assuming predeterminedness or strict exogeneity directly.

For both types of models the usual claim for a structural interpretation is that an equation or block of equations can be altered or replaced to represent a particular intervention, for example a change in policy behavior. The most important difference between their standard specifications is that SVAR's usually assert diagonality of Σ. Because SEM's do not usually assert diagonality, they leave an ambiguity as to how to interpret "changing a structural equation".

Consider the most common use of stochastic models to develop alternative policy scenarios. Usually this is done by setting aside the policy equation or block of equations and fixing a hypothetical time path for "policy variables". In monetary policy analysis this most commonly involves fixing a time path for an interest rate thought to be controlled by policy. In a standard SVAR, it is natural to suppose that this intervention leaves the joint distribution of the non-policy disturbances in the system unaffected. But in a standard SEM, the model implies that in the past disturbances to policy were correlated with disturbances elsewhere in the system. So how should we "hold constant" the distribution of the non-policy disturbances when we fix a time path for r? We could simply hold the joint distribution of non-policy disturbances constant. This would imply that we interpret the historical correlation of disturbances as reflecting causal influence of non-policy shocks on policy behavior. At the opposite extreme, we could use the hypothetical time path of r and the original policy equation to generate an implied sequence of policy equation disturbances, then use the historical correlation patterns to generate an implied time path for the vector of non-policy disturbances. This would imply that we interpret the historical correlation of disturbances as reflecting causal influence of policy shocks on non-policy behavior.

Neither of these approaches is satisfying. If we think that historical correlations of policy with non-policy disturbances reflect influences of private sector behavior

on policy, then these influences ought to be accounted for in our policy behavior equations, and vice versa if we think causality has run the other way. So long as disturbances are correlated, the model does not provide a complete description of how to model a policy change.

While this is an important difference between the actual practice of most SEM modelers and SVAR modelers, it is possible to incorporate restrictions on Σ into an SEM framework[1], and it is possible to use exogeneity and predeterminedness assumptions in an SVAR framework. The boundaries between the two types of model are indistinct.

3. Structural VAR and SEM Identification

In both forms of model we assume Γ_0 is full rank, so the **reduced form** (3) exists. This is the assumption that the model is **complete**. It implies that the system can be solved to determine $y(t)$ from past values of y and current shocks $\varepsilon(t)$. In the SVAR model and the SEM model with predetermined z's, the conditional distribution of $y(t)$ given past y's is therefore determined by the coefficients in the reduced form lag polynomial $B(L)$, by $\gamma = \Gamma_0^{-1}c$, and by the parameters of the conditional distribution of $\nu(t) = \Gamma_0^{-1}\varepsilon(t)$ given past y's. In the SVAR case, the conditional distribution of $\nu(t)$ is $N\big(0, \Gamma_0^{-1}(\Gamma_0^{-1})'\big)$. Many matrices A will satisfy $AA' = \Gamma_0^{-1}(\Gamma_0^{-1})'$. One way to see why is to note that there are only $(n+1)n/2$ free elements of $\Gamma_0^{-1}(\Gamma_0^{-1})'$, because of its symmetry, while Γ_0 itself has n^2 free elements. Thus we could never solve for Γ_0 if given the covariance matrix of $\nu(t)$, because we would have fewer equations than unknowns. But since $B_0 = I$ by construction, and otherwise the number of paraemters in $B(L)$ matches the number in $\Gamma(L)$, the reduced form as a whole has $(n-1)n/2$ fewer parameters than the standard structural form (2). Since the properties of the data are determined by the reduced form parameters, any attempt to determine the structural parameters from properties of the data will face indeterminacy, unless we can find $(n^2-n)/2$ identifying restrictions.

For an SEM model, there is a similar indeterminacy. The standard SEM implies that the only connection between Γ_0 and $\Sigma = \mathrm{Var}(\nu(t))$ is that Γ_0 must be block triangular under the same ordering of variables that makes Σ block diagonal. Usually it is assumed that the equations in the lower block, determining the z variables, are not structural, so that indeterminacy in that block is resolved by arbitrary normalizing assumptions (if that block is estimated at all). The upper block then has $\Gamma_{11}(L)$, $\Gamma_{12}(L)$, and Σ_{11} as free parameters. The corresponding part of the reduced form has fewer parameters by n_1^2, where n_1 is the number of variables in the upper block (i.e., the number of **endogenous** variables in the usual terminology). So in general we need n_1^2 a priori restrictions or normalizations to eliminate indeterminacy in the SEM model.

[1] And indeed, via exogeneity and predeterminedness assumptions, the SEM framework already does so.

Another way to see the same point is to note that if (2) represents a standard SVAR, then we can multiply it through on the left by an arbitrary orthonormal matrix Q and have a model, with parameters $\Gamma^*(L) = Q\Gamma(L)$ and $c^* = Qc$, that has the same implications for the behavior of the data and remains in the standard SVAR class. An orthonormal matrix of order n has $(n-1)n/2$ free parameters.

In a standard SEM, the upper block can be multiplied through by an arbitrary non-singular matrix, to produce a new model that is in the standard SEM form and has the same implications for the data's behavior as the original model. Such a matrix has n_1^2 free coefficients, where n_1 is the number of endogenous variables.

4. SVAR CONTEMPORANEOUS RESTRICTIONS

Much of the SVAR literature works with restrictions on Γ_0 alone. One reason for this is substantive. Such restrictions usually have an interpretation as assumptions about delays in the reaction of particular classes of agents to disturbances originating outside their own sector. While this kind of restriction does not flow from any fancy economic theory, it is relatively easy to assess and argue about based on observations of real people and institutions.

But at least as important in the popularity of this type of restriction is that it interacts conveniently with the structure of the likelihood function. The model can be thought of as parameterized in terms of $B(L)$ and Γ_0, with Γ_0 simply determining the reduced form residual covariance matrix. If disturbances are Gaussian, the log likelihood then has the form

$$\ell(B,\Gamma_0) = -\tfrac{1}{2}T\log(2\pi) + T\log|\Gamma_0| - \frac{1}{2}\operatorname{trace}\big(S(B,\gamma)\Gamma_0'\Gamma_0\big)\,,$$

where

$$u_t(B,\gamma) = B(L)y_t - \gamma\,, \qquad S(B,\gamma) = \sum_{t=1}^{T} u_t(B,\gamma)u_t(B,\gamma)'\,.$$

The maximum likelihood estimator of (B,γ) is found by equation-by-equation OLS, so long as (B,γ) is unrestricted, regardless of the value of Γ_0. This is just a version of the standard "seemingly unrelated regressions" result that when all equations in a system have the same right-hand-side variable list, equation-by-equation OLS is the MLE, regardless of the covariance matrix of residuals.

If we integrate the likelihood (or, with a conjugate prior, the posterior pdf) with respect to (B,γ), the resulting log marginal pdf for Γ_0 is proportional to

$$-\tfrac{1}{2}(T-k)\log(2\pi) + (T-k)\log|\Gamma_0| - \tfrac{1}{2}\operatorname{trace}\big(S(\hat{B},\hat{\gamma})\Gamma_0'\Gamma_0\big)\,, \tag{5}$$

where k is the number of right-hand-side variables in each regression and $\hat{B},\hat{\gamma}$ are OLS estimates. If instead we concentrate the likelihood with respect to (B,γ) (i.e., maximize it with respect to (B,γ), holding Γ_0 fixed), we get the same expression

except with T rather than $T-k$ multiplying the first term. (It is common for researchers to use $|\Gamma_0|^k$ as an improper reference prior in these models, so that concentrated likelihood and marginal posterior coincide).

The result is that to maximize the likelihood, or the marginal posterior for Γ_0, one can proceed in two steps. First apply OLS to obtain estimates of (B, γ) and construct $S(\hat{B}, \hat{\gamma})$. Then maximize (5) with respect to Γ_0. This latter step is likely (unless the model is exactly identified) to require numerical non-linear maximization, but the number of parameters to be handled is many fewer than would have to be considered jointly if the optimization were over all the parameters in $\Gamma(L)$.

This special structure of the likelihood also simplifies making draws from the posterior pdf. The marginal distribution for Γ_0 is non-standard and may therefore require MCMC methods, but the distribution of (B, γ) given Γ_0 is Gaussian.

5. Long run restrictions

Some of the SVAR literature (Blanchard and Quah, 1989, e.g.) uses what are called "long run restrictions". These are restrictions on sums of coefficients in $\Gamma^{-1}(L)$ and thus do not fit within the restrictions-on-Γ_0-only framework of the previous section. These restrictions usually are said to arise from somewhat more elegant economic theory than the "delayed reaction" theory underlying most zero restrictions on Γ_0.

For example, it might be argued that the long run effects of a monetary policy shock on M, P and W (money stock, prices, and wages, in logs) should all be the same. This might then be taken to mean that, in a system where these variables appear in differenced form, the column of $\Gamma^{-1}(1)$ corresponding to the monetary policy shock, the elements corresponding to these three variables should all be the same.[2]

Blanchard and Quah (henceforth BQ) consider a system in the two variables Δy and u, the growth in the log of output and the unemployment rate respectively, postulate that it is driven by two orthogonal structural shocks ("supply" and "demand"), and argue that a demand shock should have no permanent effect on the level of output, whereas a supply shock should.

Restrictions on $\Gamma(1)^{-1}$ generally translate into nonlinear restrictions on $\Gamma(1)$ itself and thus on the coefficients in $\Gamma(L)$, which is inconvenient. The BQ restrictions, though, because their system is so small, are not so inconvenient. They amount to requiring (if the demand shock is second) that the upper right element in their two by two $\Gamma^{-1}(1)$ matrix be zero, i.e. that that matrix be lower triangular. But triangularity is preserved under inversion, so the restriction is equivalent to requiring that $\Gamma(1)$ be lower triangular, which is a linear restriction.

[2] Here we are using the common convention that $\Gamma^{-1}(z)$, when its argument is a number or a numerical variable instead of the lag operator L, is evaluated as a polynomial on the complex plane instead of as a polynomial in the lag operator.

Most applications of long run restrictions work with exactly identified models. That is, they work with models in which the number of restrictions is exactly enough to create a one-one mapping between the restricted Γ and the reduced form parameters B, Ω, where Ω is the covariance matrix of reduced form residuals. This means that estimation is by OLS, followed by solving a set of nonlinear equations. If instead the model has more restrictions, the whole estimation process becomes a nonlinear optimization, jointly over all the coefficients in $\Gamma(L)$.

The theory underlying this type of restriction is shakier than it may seem. Neutrality restrictions, for example, have no implications for the coefficients in a stationary model. What theory is usually taken to assert, at least approximately, is that *if* random disturbances to policy produce a permanent shift in the level of the money stock, then all other nominal variables should change in proportion. But if, in the data at hand, money stock has been stationary about a deterministic trend, this theory makes no assertions about the coefficients in the representation of the data. Or, if the nominal data are non-stationary, but policy shocks themselves are not the source of the non-stationarity, then again the theory places no restrictions on the coefficients of the monetary policy shock in the impulse responses of the model.

In the BQ model, identification rests not only on the claim that demand shocks have no long run effects on the output level, but also on the claim that the other shock does have such long run effects. Conditional on the model specification, there is actually strong evidence that supply shocks have permanent effects on the level of output, as can be seen from the analysis of the model in Sims and Zha (1998).[3] However, BQ did not include a constant term in their model and made no comparison of their specification to one that had log output stationary about a linear trend. Other researchers have found it hard to distinguish a trend-stationary from a unit-root model for log output in the US. It is therefore arguable that the identifying assumptions in the model are dubious, even if the model is accepted as correct.

An even stronger criticism is that few believe that in real economies there are just two, orthogonal, "aggregate demand" and "aggregate supply", behavioral shocks. If, for example, there were two types of technology shocks ("weather" and "science", e.g.), only one of which produced non-stationary effects on output, then the BQ identification scheme would fail.

6. SEM'S IN PRACTICE

The description here of SEM's applies to an idealized version of such models that scarcely exists today. Large models that are descendants of SEM's do exist and play an important role in the policy process. The US Federal Reserve Board, the ECB, the British NIESR, the commercial groups DRI and Wharton Econometrics, all maintain descendants of SEM's. However these models are in practice cut free from the

[3] The error bands shown in the original BQ article are incorrect, and thus can't be used to consider this kind of question.

SEM statistical theory that developed in the 1950's. They are estimated equation by equation for the most part, with no consideration of the joint likelihood of all the variables the model implies need to be considered simultaneously. They also include "forward-looking" terms — variables in the form $E_t X_{t+1}$ — which are treated more or less carefully in model simulation but are not integrated into a multivariate inference framework.

It is an interesting question why these models retain their appeal to policy-making institutions and how they have sustained some credibility even as their claims to being probability models of the data have been abandoned.

7. Identification via Impulse Responses

In the applied SVAR literature the explicit identifying restrictions on Γ_0 or $\Gamma(1)$ are not the only information that is in fact used in identification. Researchers experiment with the model specification until results start to look "reasonable". In the monetary policy SVAR literature, reasonable behavior is usually taken to mean that monetary contraction should at least on impact raise interest rates, at least in the long run lower prices and money stock, and lower, or least not increase, output.

The search for models that produce such results is actually part of the estimation process, and the "reasonableness" criteria are actually identifying restrictions. It is a recognized defect of the literature that it does not handle this identifying information formally, whether by imposing deterministic restrictions or using priors that express the belief in reasonableness.

The ideal solution to this problem would be to incorporate a strong belief in these properties into a prior on the model parameters. But the mapping between parameters and impulse responses in the 6-10 variable models that dominate this literature is so nonlinear and complicated that there are apparently no published papers that have actually implemented such a prior. There are a few papers that have made more formal use of this kind of identifying information in other ways (Uhlig, 2001; Faust, 1998, e.g.). It seems that if priors were introduced in the form of "dummy observations" (i.e., multiplicative factors) on linear functionals of the impulse responses, an internally consistent Bayesian analysis incorporating "reasonableness priors" should be possible, but it hasn't been done yet.

8. The Lucas Critique

Use of SVAR's to analyze the effects of variations in monetary policy is sometimes taken to be "subject to the Lucas Critique". The Lucas critique observed that SEM-style models that assume that private agents' expectations are fixed linear functions of lagged data are likely to be mistaken in projecting the effects of systematic changes in monetary policy. Such a policy change would be likely to change the optimal forecasting formula, and thus to change the dynamics of private sector behavior, according to the SEM models themselves.

The interventions a monetary policy SVAR is designed to analyze are precisely the sort of thing Lucas warned could lead to inconsistencies — changes in monetary policy behavior equations. However, SVAR's, unlike the old SEM's, do not contain fixed-coefficient expectational rules. They are best thought of as giving linear approximations to the behavior of the private sector and monetary authorities. The private behavior they model thus implicitly includes dynamics arising from revision in forecasting rules as well as other sources of dynamics.

Suppose for example that policy alternates at random but fairly long intervals between two distinct linear behavioral rules for interest-rate setting. The private sector will therefore constantly be assessing the history of interest rate changes, trying to decide which rule is currently in effect. There will be some local linear approximation to the actual nonlinear behavioral rule, and disturbances from that linear approximation will have effects both directly and indirectly through their effects on the public's assessment of the probabilities of the two regimes. An approximate linear SVAR may do quite well in projecting the effects of its identified monetary policy shocks, so long as the model's nonlinearity is not too severe. A setup like this was modeled by Cooley, Leroy, and Raymon (1984). The same arguments would apply even if the policy regimes jumped in a non-stationary way from one linear rule to another instead of varying over a given finite set.

Of course if policy is jumping between linear rules and the public is trying to assess when the rule changes and how, the entire model will be nonlinear, so that linear approximations could be inaccurate. How inaccurate they are will depend on how great the nonlinearity is and on exactly what sequence of shocks is fed into the model. If the model appears to fit historical data well and shows little sign of nonlinearity in the sample period, then policy changes that produce policy equation disturbances in patterns similar to what has been observed in the past are likely to be projected accurately by the model, even if they have been generated by a "change in rule", in the sense of a change in the coefficients in a linear policy behavior equation.

Issues of this type are discussed at more length in Leeper and Zha (2001) and Sims (1987).

REFERENCES

BLANCHARD, O., AND D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, 655–73.

COOLEY, T. F., S. F. LEROY, AND N. RAYMON (1984): "Econometric Policy Evaluation: Note," *American Economic Review*, 74, 467–70.

FAUST, J. (1998): "The Robustness of Identified VAR Conclusions About Money," *Journal of Monetary Economics*, 49, 207–244, Carnegie-Rochester Conference Series on Public Policy.

HURWICZ, L. (1962): "On the Structural Form of Interdependent Systems," in *Logic, Methodology and Philosophy of Science*, pp. 232–239. Stanford University Press, Stanford, CA.

LEEPER, E., AND T. ZHA (2001): "Modest Policy Interventions," Discussion paper, Indiana University and Federal Reserve Bank of Atlanta, `http://php.indiana.edu/~eleeper/Papers/lz0101Rev.pdf`.

SIMS, C. A. (1987): "A Rational Expectations Framework for Short Run Policy Analysis," in *New Approaches to Monetary Economics*, ed. by W. Barnett, and K. Singleton, pp. 293–310. Cambridge University Press.

SIMS, C. A., AND T. ZHA (1998): "Error Bands for Impulse Responses," *Econometrica*, 67, 1113–1156.

UHLIG, H. (2001): "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure," Discussion paper, Humboldt University, Berlin.