18 Nonlinear Structural VAR Models

18.1 Motivation

The standard VAR model, as discussed in previous chapters, is designed to capture the linear dependence of y_t on its own lags. This model is linear in the slope parameters as well as linear in the lagged model variables. More generally, however, the conditional mean may be nonlinear in the lagged variables and/or the model parameters.

Consider the class of VAR models where y_t depends nonlinearly on its lags. A K-dimensional nonlinear VAR process for y_t may be defined as

$$y_t = \mathbf{F}_t(y_{t-1}, \ldots, y_{t-p}) + u_t, \tag{18.1.1}$$

where the white noise reduced-form innovations are additively separable and the nonlinear function $\mathbf{F}_t(\cdot)$ may depend on t. The appropriate form of the function $\mathbf{F}_t(\cdot)$ depends on the economic context.

If $\mathbf{F}_t(\cdot) = \mathbf{F}(\cdot)$, as long as the function $\mathbf{F}(\cdot)$ is well behaved, the nonlinear conditional mean specification may equivalently be expressed as

$$\begin{aligned} y_t = {} & \nu + \text{linear component} + \text{quadratic component} \\ & + \text{cubic component} + \cdots + u_t, \end{aligned}$$

based on a Taylor-series expansion of the conditional mean about zero. For example, in a model including only one autoregressive lag, the quadratic component would involve all squares and pairwise products of the elements of y_{t-1}, and the cubic component would contain regressors of the form $y_{m,t-1}y_{n,t-1}y_{l,t-1}$, with m, n, and $l \in \{1, \ldots, K\}$. In that case, the standard linear reduced-form VAR model

$$y_t = \nu + A_1 y_{t-1} + u_t,$$

may be viewed as a first-order linear approximation to this more general process.

Often the linear VAR approximation is quite accurate, but it is important to recognize that the fact that the approximating model is linear in the lagged variables and linear in the autoregressive slope parameters imposes several restrictions on the structural impulse response functions. First, the impulse responses increase proportionately with the magnitude of the structural shock. Second, responses are symmetric in positive and negative structural shocks. In other words, the response to a positive structural shock of given magnitude is the exact mirror image of the response to a negative structural shock of the same magnitude. Third, the responses to structural shocks are invariant to when the structural shock occurs and to the state of the economy, at the time when the structural shock occurs. There are situations in which these restrictions are clearly unrealistic, invalidating the use of linear approximations.

For example, in international trade transportation costs may prevent arbitrage between the prices of the same good in different countries expressed in the same currency. Arbitrage will occur only when the price differential exceeds the transportation cost. The existence of such thresholds implies that for small values of the price differential, there will be no response to price shocks, whereas for higher values there will be a disproportionate adjustment, invalidating the assumption of linearity. Similar zones of inaction arise in financial economics when trades involve a transaction cost. Other potential sources of thresholds include capacity constraints in production as well as ceilings and floors in modeling inventories. In addition, thresholds may arise from government regulation. Exchange rate target zones are a good example. It is possible to adapt the linear VAR model to allow for such threshold dynamics, resulting in a threshold VAR model.

Often the rationale for hard thresholds of this type is compelling at the microeconomic level, but not at the macroeconomic level. Aggregation across firms or households with different thresholds implies the existence of smooth thresholds (or smooth transitions) in the data. Thus, when working with aggregate data, it is common to postulate a smooth transition function that depends on how far the variable of interest is from its latent equilibrium value. Typical choices include an exponential or a logistic function.

Such smooth-transition models are also useful in modeling traders' uncertainty about the correct specification of their model as well as, more generally, heterogeneity in beliefs among traders. For example, even if a given trader were convinced that a dollar-euro exchange rate at parity is the equilibrium value, there is enough disagreement among traders and enough uncertainty about the equilibrium exchange rate that it would be unwise for that trader to take a position against an exchange rate that is only slightly higher than parity. As a result, arbitrage does not take place, and the exchange rate may evolve seemingly at random in the neighborhood of the latent equilibrium. A market consensus that the exchange rate is overvalued only emerges if the exchange rate reaches much higher values. Once there is a consensus that the exchange rate is overvalued, there will be a disproportionate increase in traders anticipating a decline in the exchange rate and acting on these beliefs, ensuring that the exchange rate ultimately reverts back toward the latent equilibrium. These corrective forces, however, will weaken, as the exchange rate approaches the latent equilibrium. Smooth-transition VAR models are well suited to capturing such phenomena. It is also possible to adapt VAR models to allow for smooth one-time transitions following major discrete structural changes. For example, one may model the transition of inflation to a new steady state following the introduction of a new monetary system or a shift in the credibility of the central bank.

Another important type of nonlinearity involves the economy alternating between two or more states (or regimes), say the economy being in recession and being in expansion, with the VAR model parameters differing across regimes and with the transition between the regimes being governed by a stochastic process. Such regime-switching models allow the effect of structural shocks to differ across regimes. For example, fiscal policy shocks may have larger effects during a recession than during an expansion. This notion plays an important role in the debate about the magnitude of the fiscal multiplier in macroeconomics. More generally, even the sign of the response may be affected by the state of the system at the time of the shock.

It is also possible to allow measures of uncertainty to have a direct effect on the conditional mean of economic time series. Such a specification can be motivated by real options theory, which implies that uncertainty about the price of a commodity may delay investment decisions, if the cash flow of the investment depends on that commodity price. For example, it has been argued that shifts in the conditional variance of the percent change in the real price of oil lower U.S. real GDP growth. Such nonlinearities may be captured using GARCH-in-mean VAR models or VAR models with stochastic volatility in the error term.

Another potential source of nonlinearities is smooth structural change. Smooth structural change calls for the use of time-varying coefficient VAR (TVC-VAR) models that allow the VAR model coefficients to evolve continuously over time, according to a prespecified law of motion. Such models are designed for situations when the data are subject to many potential sources of nonlinearities of unknown form. For example, in modeling the global market for crude oil, changes in market structure, changes in energy efficiency and energy conservation, temporary capacity constraints, the emergence of unconventional oil production, delays in the response of new oil production to price incentives, gradual substitution among different forms of energy consumption, changes in refining and transportation infrastructure, and regulatory changes, all may render the VAR model coefficients unstable over time. If this instability is quantitatively important, a TVC-VAR model may be preferable to a linear VAR model.

Finally, the linear VAR model has been adapted to allow for responses that are asymmetric in positive and negative structural shocks. Such asymmetries are often modeled by censoring some of the model variables. This practice complicates the identification of the structural shocks and invalidates standard methods of estimation and inference based on linear VAR models, even when the model remains linear in the parameters. There are alternative model specifications, however, that remain valid, whether the responses in the DGP are symmetric or not.

As this overview shows, there are many types of nonlinear VAR models, each of which calls for a different specification of the function $\mathbf{F}_t(\cdot)$. The common feature of all these models is that they allow current values of a set of variables to depend nonlinearly on lagged values of these variables. Referring to such models as nonlinear VAR models is conventional, but technically misleading because genuine VAR models are linear. We, nevertheless, follow this convention and broaden our definition of VAR models in this chapter to include nonlinear specifications.

The focus of the chapter is on the use of nonlinear VAR models for structural analysis. A detailed account of the specification and estimation of nonlinear time series models, including many of those relevant in the current context, is provided in Granger and Teräsvirta (1993) and Teräsvirta, Tjøstheim, and Granger (2010).

Section 18.2 outlines the general setup of nonlinear VAR models and reviews some of the challenges involved in conducting structural analysis in nonlinear models. In the subsequent sections a number of special cases of nonlinear structural VAR models are considered. We focus on threshold and smooth-transition VAR models (Section 18.3), Markov-switching VAR models (Section 18.4), time-varying coefficient VAR models (Section 18.5), and GARCH-in-mean VAR models (Section 18.6). A brief review of nonparametric and semiparametric VAR models is presented in Section 18.7. Some general comments on nonlinear VAR modeling are provided in Section 18.8. Finally, Section 18.9 focuses on models that are linear in the parameters, but are nonlinear in the variables.

18.2 Nonlinear VAR Analysis

18.2.1 General Setup

The standard linear VAR model has the reduced form

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

where y_t is a K-dimensional vector of observed variables, ν is a $K \times 1$ constant term, $A_1, \ldots, A_p$ are the $K \times K$ VAR slope coefficients, and u_t is an unobserved white noise error term. Thus, the current vector of observations, y_t, depends linearly on the past values of the variables under consideration. The corresponding structural form is

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

where w_t is the structural error term with diagonal covariance matrix. The structural form allows us to explicitly model the instantaneous relations between the variables. The structural errors are linearly related to the variables. Their impact effects are given by the matrix B_0^{-1}.

This model may be generalized to allow for nonlinearities by considering structural processes of the form

$$\mathbf{G}_t(y_t, y_{t-1}, \ldots, y_{t-p}) = w_t, \tag{18.2.1}$$

where the structural shocks w_t may have nonlinear impact effects on y_t, and $\mathbf{G}_t(\cdot)$ is a nonlinear function of current and lagged data that depends on t (see Mann and Wald 1943). The reduced form of model (18.2.1) may be obtained by expressing y_t as a nonlinear function of the lagged variables.

Much of the literature on nonlinear VAR models focuses on models with the somewhat simpler reduced-form representation (18.1.1),

$$y_t = \mathbf{F}_t(y_{t-1}, \ldots, y_{t-p}) + u_t, \tag{18.2.2}$$

where the white noise reduced-form innovations are additively separable and the nonlinear function $\mathbf{F}_t(\cdot)$ may depend on t. Because the error term enters linearly, $u_t = B_0^{-1} w_t$ as in the linear model. There is, of course, no compelling economic reason why structural shocks in general should not be nonlinear functions of the reduced-form errors. In the latter case, one would have to estimate the structural model (18.2.1) directly rather than estimating the reduced form first and then recovering the structural model parameters from the reduced-form model.

It is important to recognize that the presence of nonlinearities in the DGP implies time variation in the coefficients of the linear VAR representation. Some of the nonlinear models considered in this chapter can be viewed as special cases of the time-varying coefficient model

$$y_t = \nu(t) + A_1(t) y_{t-1} + \cdots + A_p(t) y_{t-p} + u_t, \tag{18.2.3}$$

where u_t is white noise and where the dependence of the coefficients on t indicates that the slope coefficients $\boldsymbol{\alpha}(t) = \text{vec}[A_1(t), \ldots, A_p(t)]$ evolve according to a linear VAR(1) process

$$\boldsymbol{\alpha}(t) = \Gamma \boldsymbol{\alpha}(t-1) + \eta_t, \tag{18.2.4}$$

with Γ denoting a $pK^2 \times pK^2$ transition matrix consisting of time-invariant parameters and the error η_t following a pK^2-dimensional white noise process that is independent of u_t. Such models are sometimes referred to as time-varying parameter VAR models in the literature. A more appropriate terminology is time-varying coefficient VAR (TVC-VAR) models because parameters in frequentist analysis are time-invariant.

Time-varying coefficient VAR models can capture very general nonlinear dynamics. They can also accommodate discrete structural changes, provided the specification of model (18.1.1) is chosen appropriately. For example, if

$$[\nu(t), A_1(t), \ldots, A_p(t)] \\ = \begin{cases} [\nu_1, A_{1,1}, \ldots, A_{p,1}] & \text{for } t = 1, \ldots, T_B, \\ [\nu_2, A_{1,2}, \ldots, A_{p,2}] & \text{for } t = T_B + 1, \ldots, T, \end{cases}$$

the change in coefficients is best understood as a discrete structural shift in the model coefficients after time T_B. Clearly, this process is nonstationary. Likewise, time-varying coefficient VAR models can be used to capture smooth structural changes in the model coefficients by postulating that the elements of $\boldsymbol{\alpha}(t)$ follow independent random walks, for example.

Alternatively, a time-varying coefficient VAR model of the form (18.1.1) may be used to capture genuine nonlinearities that arise in the absence of structural change. In that case, the DGP may in fact be stationary, allowing one to construct the MA representation of the data,

$$y_t = \mu_y + v_t + \sum_{i=1}^{\infty} \Phi_i v_{t-i}. \tag{18.2.5}$$

Here v_t is a white noise process, and v_t and v_s are uncorrelated, but not necessarily independent for $t \neq s$. The MA representation (18.2.5) is of limited use in practice, however, because it obscures the nonlinear relations between the observed variables. These features are hidden in the potentially complex dependence structure of v_t. As a result, the MA representation of stationary nonlinear models cannot be used for constructing statistics such as impulse responses and hence is irrelevant for the analysis in this chapter.

18.2.2 Structural Analysis

One of the drawbacks of working with nonlinear structural models is that in practice the analysis is restricted to impulse response analysis and to assessing the incremental effect of a sequence of structural shocks. In contrast, it is not obvious how to define forecast error variance decompositions in nonlinear structural VAR models nor is it clear how to construct historical decompositions. Our discussion in this section focuses on structural impulse response analysis. It is useful to first outline the general principles in the context of model (18.1.1). For expository purposes, it is assumed for now that B_0^{-1} is known, allowing us to infer the structural shocks from the reduced-form shocks.

The nonlinearity of the conditional mean specification complicates the construction of structural impulse responses. Because impulse responses depend on the sign and magnitude of the structural shock as well as on the history of the data, the standard approach to constructing structural impulse responses discussed in Chapter 4 cannot be used when working with nonlinear VAR models.

Structural impulse responses measure by how much future realizations of y_{t+h} for $h = 0, 1, \ldots, H$ are expected to differ in response to a one-time structural shock at date t, compared with a baseline in which no such shock occurs. A natural definition of a nonlinear structural impulse response thus is as the difference between two conditional expectations of the realizations of y_{t+h}, $h = 0, 1, 2, \ldots, H$, where the first expectation is conditional on the information set available at date $t-1$, denoted Ω_{t-1}, as well as on the magnitude δ of the i^{th} structural shock, whereas the second expectation only conditions on Ω_{t-1}, but not on δ. This structural impulse response is a conditional impulse response in that its definition depends on the particular history Ω_{t-1}. More formally, we can define the conditional response function with respect to the i^{th} structural shock up to horizon H as the $K(H+1)$-dimensional vector

$$I_y(H, \delta, \Omega_{t-1}) \equiv \begin{pmatrix} \mathbb{E}(y_{t+0}|w_{it} = \delta, \Omega_{t-1}) - \mathbb{E}(y_{t+0}|\Omega_{t-1}) \\ \vdots \\ \mathbb{E}(y_{t+H}|w_{it} = \delta, \Omega_{t-1}) - \mathbb{E}(y_{t+H}|\Omega_{t-1}) \end{pmatrix}, \tag{18.2.6}$$

where Ω_{t-1} consists of the history of the model data up to time $t-1$. How many lagged values of y_t are included in this history depends on the lag structure of the underlying model.

In practice, the expectations in question must be evaluated by Monte Carlo integration. Having estimated the K-dimensional nonlinear reduced-form model (18.1.1) and having recovered the sequence of structural shocks in the structural representation (18.2.1), the structural impulse responses may be computed by the following bootstrap procedure, if we are willing to impose the additional assumption that the structural shocks are mutually independent rather than merely mutually uncorrelated. Under the added assumption of Gaussian errors, these conditions would be equivalent.

Algorithm *Conditional impulse response function at date t*

1. The sequence of lagged data up to period $t-1$ defines the history Ω_{t-1} at date t.
2. Consider a structural shock of magnitude δ at date t in one of the elements of w_t. Given Ω_{t-1}, we simulate two time paths of the realizations of y_{t+h}, $h = 0, \ldots, H$. When generating the first time path,

the value of the structural shock of interest is set equal to the prespecified value δ, denoting the magnitude of this shock. For example, we might set δ equal to 1. Subsequent realizations of this shock for period $t + h$, $h = 1, \dots, H$, are drawn from the marginal empirical distribution of the structural shock of interest. The realizations of the other $K - 1$ structural shocks for $h = 0, 1, \dots, H$ are all drawn independently from their respective marginal distributions. Given this K-variate sequence of structural shocks, we simulate the path of y_{t+h}, $h = 0, \dots, H$, by recursively updating the fitted model (18.1.1), conditional on Ω_{t-1}.

This simulated path of the data must be compared to the baseline path in which no structural shock of magnitude δ occurs in period t. When generating this second time path, all K structural shocks for $h = 0, \dots, H$ are drawn independently from their respective estimated marginal distributions. Given this alternative K-variate sequence of structural shocks, we again simulate the path of y_{t+h}, $h = 0, \dots, H$, by recursively updating the fitted model (18.1.1), conditional on Ω_{t-1}.

3. We now subtract the second time path for y_{t+h}, $h = 0, 1, \dots, H$, from the first time path. This difference is an estimate of the structural impulse response function conditional on Ω_{t-1}, but this estimate is noisy because it depends on random draws for the structural shocks, and hence is of little use in practice.
4. To eliminate the random variation in the impulse response estimate we repeat steps 2 and 3 many times and average the resulting impulse response estimates. By the law of large numbers, this average converges to the conditional response of y_{t+h} at horizon $h = 0, 1, \dots, H$ to a shock of magnitude δ conditional on Ω_{t-1}:

$$I_y(H, \delta, \Omega_{t-1}).$$

This conditional response function is of interest, if we are concerned with the response of the model variables at a particular point in time. An economist studying the performance of the economy over a longer time span based on a stationary nonlinear model, in contrast, may be interested in the unconditional impulse response function instead. The unconditional response of the variable y is defined as

$$I_y(H, \delta) = \int I_y(H, \delta, \Omega^r) d\Omega^r,$$

where Ω^r is a randomly selected history. In practice, $I_y(H, \delta)$ is simply obtained by averaging the value of the conditional response function $I_y(H, \delta, \Omega^r)$ over many histories, Ω^r, each of which is randomly drawn with replacement from the original data.

Alternatively, one could also condition on the state of the economy. For example, one might condition on the average of the subset of all histories in which the economy is in recession.

Regardless of whether we condition on a particular history or not, an important question is how to choose δ. One possibility is to report the range of nonlinear structural responses for a grid of δ values, keeping in mind that the magnitude of δ must not exceed the support of the structural residuals. It is also common to focus on one standard deviation and two standard deviations of the structural shock of interest, representing a shock of typical magnitude and an unusually large (but not unrealistically large) shock, respectively. Note that in comparing or averaging conditional impulse response functions across time it is important to keep constant the value of δ.

Impulse responses constructed in this manner are sometimes referred to as generalized impulse responses. Generalized impulse response functions were first proposed by Koop, Pesaran, and Potter (1996) who laid the foundations for nonlinear VAR impulse response analysis.[1] It is important to note, however, that Koop et al.'s analysis was not concerned with structural impulse response analysis, but with reduced-form impulse response analysis. Their focus was on variable-specific or system-wide reduced-form shocks drawn from the joint distribution of u_t. In contrast, our focus in this chapter is on characterizing nonlinear responses to innovations in w_t. Hence, the algorithm described above differs from Koop et al.'s original proposal.

There is some confusion in the literature about the precise sense in which Koop et al.'s analysis generalizes the traditional approach to impulse response analysis. As a result, the term generalized impulse response has been used to denote very different types of impulse responses in applied work.

Without doubt, the definition of the impulse response function as the difference between two conditional expectations is more general than the definition in Chapter 4 in that it can be applied to both linear and nonlinear VAR models. Moreover, it can be shown that this definition reduces to the standard definition of structural impulse responses when the VAR model is linear.

This is not the only sense in which definition (18.2.6) has been considered a generalization of the traditional approach, however. For example, Pesaran and Shin (1998) explicitly advocate the use of the generalized impulse responses proposed in Koop, Pesaran, and Potter (1996) as being more general than traditional structural VAR impulse responses even in linear models. They suggest that studying the response of the model variables to reduced-form shocks is more general because it avoids the use of economic identifying assumptions that may be controversial. This argument is not persuasive because

[1] A related approach is the conditional moment profile of Gallant, Rossi, and Tauchen (1993) who consider not only the conditional mean as in (18.2.6) but also higher conditional moments such as the conditional variance.

reduced-form shocks have no economic interpretation and violate the ceteris paribus assumption required for quantifying causal relationships in the data. At best, these responses may serve as a description of the properties of the data. They are not informative about the structural model. Moreover, if we were actually interested in responses to reduced-form shocks in linear VAR models, such responses could easily have been generated using the standard tools already discussed in Chapter 4.

In light of these conflicting interpretations, the terminology of generalized impulse responses should perhaps be avoided. Indeed, it is not uncommon in applied work for researchers not to be explicit about how their generalized responses are constructed or to give the misleading impression that structural responses may be obtained from reduced-form models without further identifying assumptions. When applying definition (18.2.6) or its generalizations to structural nonlinear VAR models in this chapter, we therefore simply refer to nonlinear structural impulse responses.

So far we have maintained the simplifying assumption that B_0^{-1} is known, allowing us to map the reduced-form innovations into structural shocks. In practice, B_0^{-1} is not known, of course, but has to be estimated from the data. Like in linear VAR models the challenge is how to identify the structural shocks. As long as we restrict attention to models of the form (18.1.1), this question may be addressed in much the same way as in linear models because the structural shocks are linear transformations of the reduced-form residuals. For example, we may impose short-run identifying restrictions as discussed in Chapter 8.

In contrast, the use of long-run exclusion restrictions as in Chapter 10 is problematic in the current context because there is no closed-form solution for the long-run impact of structural shocks in nonlinear VAR models. Sometimes in applied work based on TVC-VAR models or other nonlinear VAR models long-run identifying restrictions are imposed on the long-run response of the instantaneously linear model observed at date t. This practice is problematic because it ignores the expected change in the model coefficients.

Likewise, identification by sign restrictions as discussed in Chapter 13 is not straightforward in the presence of nonlinearities. Typically, researchers using nonlinear sign-identified VAR models report the posterior median response function, but this practice is problematic for the reasons discussed in Chapter 13. Alternative procedures for summarizing the posterior of the structural models, as proposed in Inoue and Kilian (2013), rely on the existence of a one-to-one closed-form mapping between the VAR model parameters and the set of structural impulse responses and hence do not apply in the nonlinear setting. This problem remains unresolved thus far.

Finally, in previous chapters we have seen that the statistical properties of the model can also be helpful in structural modeling (see Chapter 16). This approach may also be feasible in nonlinear models. Although such purely statistical approaches cannot substitute for economic identifying assumptions, they may be helpful in assessing the consistency of conventional economic identifying restrictions with the data.

The following sections discuss several specific structural nonlinear VAR models and their economic applications.

18.3 Threshold and Smooth-Transition VAR Models

Threshold models allow the model coefficients to evolve from one regime to another when some model variable exceeds a prespecified threshold value. For example, the central bank may tighten monetary policy only if the inflation rate exceeds a certain level. In a smooth-transition model this transition from one regime to another is allowed to be gradual. The framework discussed in this section encompasses both of these models as special cases. Threshold and smooth-transition VAR models have been employed by Balke and Fomby (1997), Teräsvirta, Tjøstheim, and Granger (2010), Rothman, van Dijk, and Franses (2001), Camacho (2004), and Galvao and Marcellino (2014), among others. A recent survey of this literature is Hubrich and Teräsvirta (2013).

18.3.1 Model Setup

Consider the reduced-form model

$$y_t = \nu + \sum_{j=1}^{p} A_j y_{t-j} + G(x_t, \theta)\left(\nu^+ + \sum_{j=1}^{p} A_j^+ y_{t-j}\right) + u_t, \qquad (18.3.1)$$

where ν^+ is a constant. The $K \times K$ matrix function $G(x_t, \theta)$ depends on the variable x_t and the parameter θ and determines when and to what extent there is a change in the model coefficients. If $G(x_t, \theta) = 0$, the model (18.3.1) is a standard linear reduced-form VAR model with intercept term ν, slope parameter matrices A_j, $j = 1, \ldots, p$, and white noise error term u_t. If $G(x_t, \theta)$ is nonzero, the parameters ν^+ and A_j^+, $j = 1, \ldots, p$, affect the determination of y_t. The specific form of $G(x_t, \theta)$ determines the nature of the nonlinear behavior. The variable x_t may consist of lagged values of y_t, y_{t-d}, where d is known as the delay, or of additional variables not included in y_t. The precise choice depends on the economic context. A common choice in applied work is the squared deviation of some model variable from its long-run equilibrium value, as a measure of the deviation from equilibrium. Often this variable is lagged by more than one period and/or averaged over several periods. Another common choice is the cumulative change in a model variable over the d most recent periods.

For example, in an exponential smooth-transition VAR (EST-VAR) model the function $G(x_t, \theta)$ is a diagonal matrix with exponential transition functions on the diagonal such that

$$G(x_t,\theta) = \begin{bmatrix} 1-\exp[-\gamma(x_{1t}-c_1)^2] & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1-\exp[-\gamma(x_{Kt}-c_K)^2] \end{bmatrix}, \tag{18.3.2}$$

where $\gamma > 0$, $x_t = (x_{1t}, \dots, x_{Kt})'$ and $\theta = (\gamma, c_1, \dots, c_K)'$. As long as the transition variable x_{kt} is equal to the constant c_k, $1 - \exp[-\gamma(x_{kt} - c_k)^2] = 0$. If the difference between x_{kt} and c_k is very large, $1 - \exp[-\gamma(x_{kt} - c_k)^2] \approx 1$. Thus, if all transition variables x_{kt} are equal to c_k until some period T_B and then increasingly deviate from c_k, the process (18.3.1) gradually transforms itself from a VAR with coefficients $\nu, A_1, \dots, A_p$ to a VAR with coefficients $\nu + \nu^+, A_1 + A_1^+, \dots, A_p + A_p^+$, which is why this process is referred to as a smooth-transition model.

A possible shift variable is, for example,

$$x_{kt} = \begin{cases} c_k & \text{for } t < T_B, \\ t & \text{for } t \geq T_B, \end{cases}$$

where $1 < T_B < T$. The parameter γ determines the speed of the transition from one regime to the other. A very large value of γ makes for very fast changes from one regime to the other. In that case, the model is similar to the fixed threshold model discussed later.

Perhaps the most common specification involves an exponential transition function, $G(x_t, \theta) = (1 - \exp[-\gamma(x_t - c)^2])I_K$ (see, e.g., Rothman, van Dijk, and Franses 2001). More general versions of the model (18.3.1) may involve different γ parameters in different equations or additional transition functions. Moreover, the innovation variance may be allowed to change as discussed in Chapter 14. If integrated and cointegrated variables are involved, one may want to specify the ST-VAR model as a vector error correction model rather than a VAR in levels or in differences.

In contrast, if the function $G(x_t, \theta)$ has the form

$$G(x_t, \theta) = \mathbb{I}(x_t > c)I_K,$$

where x_t is a scalar variable, $\mathbb{I}(\cdot)$ is an indicator function, and c is a constant, then model (18.3.1) is called a threshold VAR (TVAR) model. The parameters change if the threshold variable x_t exceeds the value c. The value of the threshold may be known or unknown, depending on the economic context. Such models were proposed by Balke and Fomby (1997) and used by Galvao and Marcellino (2014). In threshold models it may make sense to consider a number of separate threshold values because different transmission mechanisms may be valid for very low, very large, and intermediate values of a threshold variable. More generally, one could consider a TVAR model with M regimes

$$y_t = \begin{cases} \nu^{(1)} + \sum_{j=1}^{p} A_j^{(1)} y_{t-j} + u_t^{(1)} & \text{if } x_t \le c_1, \\ \nu^{(2)} + \sum_{j=1}^{p} A_j^{(2)} y_{t-j} + u_t^{(2)} & \text{if } c_1 < x_t \le c_2, \\ \vdots & \\ \nu^{(M)} + \sum_{j=1}^{p} A_j^{(M)} y_{t-j} + u_t^{(M)} & \text{if } x_t > c_{M-1}, \end{cases}$$

where $c_1, \ldots, c_{M-1}$ are the (possibly unknown) threshold values.

The parameters of threshold VAR and smooth-transition VAR models can be estimated by nonlinear least squares (NLS) or by ML methods under standard assumptions. Estimation can be difficult if the models are large or if a specific regime does not appear very often in the sample. In addition, the ML estimator of the parameters in the transition function may not have a standard asymptotic distribution. A case in point is the ML estimator of the threshold parameter in a TVAR model (see Chan (1993) and Hansen (1997b, 2000) for discussions of asymptotic properties of ML estimators of univariate threshold AR models). Smooth-transition VAR models may also be estimated by Bayesian methods (see Gefang and Strachan 2010; Gefang 2012).

18.3.2 Example: A TVAR Model of U.S. Monetary Policy

Galvao and Marcellino (2014) construct a TVAR model for a three-dimensional system of U.S. economic variables based on quarterly data for 1960 to 2008. Let $y_t = (gdp_t, p_t, i_t)'$, where gdp_t is the log of real GDP, p_t is the log price index, and i_t is the federal funds rate. The model is partially identified. It relies on a recursive identification scheme in which the first shock has an instantaneous impact on all model variables, whereas the last shock has a nonzero impact effect only on the interest rate i_t and, hence, is interpreted as a monetary policy shock (see Chapter 8).

Galvao and Marcellino consider a TVAR model with three regimes of the form

$$y_t = \begin{cases} \nu^{(1)} + \sum_{j=1}^{p} A_j^{(1)} y_{t-j} + u_t^{(1)} & \text{if } x_t \le c_1, \\ \nu^{(2)} + \sum_{j=1}^{p} A_j^{(2)} y_{t-j} + u_t^{(2)} & \text{if } c_1 < x_t \le c_2, \\ \nu^{(3)} + \sum_{j=1}^{p} A_j^{(3)} y_{t-j} + u_t^{(3)} & \text{if } x_t > c_2, \end{cases}$$

where c_1 and c_2 are unknown threshold values. They postulate that $u_t^{(i)} \sim \mathcal{N}(0, \Sigma_u^{(i)})$, $i = 1, 2, 3$, is a normally distributed reduced-form error term whose variance-covariance matrix may depend on the regime.

The transition variable x_t is based on the deviation from a Taylor rule in period $t-1$. More precisely, the transition variable is defined as

$$x_t = 1 + 1.5(p_{t-1} - p_{t-5}) + 0.5(gdp_{t-1} - gdp_{t-5}) - i_{t-1}$$

such that x_t depends on lagged year-on-year inflation $(p_t - p_{t-4})$ and output growth $(gdp_t - gdp_{t-4})$. The Taylor rule involves annual output growth and inflation rather than the level of output and prices because the transition variable has to be stationary. A trending transition variable eventually would drive the system permanently into the first or last regime, ruling out future regime changes. Note that the transition variable is positive when the actual interest rate is below the value predicted by the Taylor rule, but is negative when i_t is larger than the rate implied by the Taylor rule such that large values of the transition variable indicate a loose monetary policy, whereas small values of x_t indicate a tight policy.[2]

Galvao and Marcellino (2014) estimate the model parameters including the threshold values c_1 and c_2 by Gaussian ML. They use information criteria for choosing between different model specifications. The structural shocks are obtained by applying a lower-triangular Cholesky decomposition to the reduced-form residual covariance matrix in each regime. The nonlinear structural impulse responses are constructed as discussed in Section 18.2.2.

Confidence bands for the nonlinear structural impulse responses are constructed by applying a residual-based bootstrap method to the estimated nonlinear model. This bootstrap method is intended to take into account the sampling uncertainty about the threshold parameters. It is not clear, however, what the theoretical justification for this bootstrap method is. In general, little is known about the asymptotic validity of bootstrapping nonlinear VAR models and the bootstrap methods employed in applied work tend to be ad hoc.

The study of Galvao and Marcellino (2014) is one example of the use of threshold VAR models. Other applications of smooth-transition and threshold models include Weise (1999), Balke (2000), Atanasova (2003), and Dios Tena and Tremayne (2009).

18.4 Markov-Switching VAR Models

Changes in regimes may also be modeled by making them dependent on a discrete Markov process with, say, M states. This approach has several advantages over the nonlinear models considered so far. In each period the state of the process is determined endogenously and the specific state can change from period to period. The process can even be in-between two states in a particular period. Thereby very flexible changes in the VAR model coefficients can be accommodated.

The theory and practice of Markov-switching VAR (MS-VAR) models was laid out by Krolzig (1997) who generalized the univariate model proposed

[2] Galvao and Marcellino (2014) also consider a generalization of this baseline TVAR model that allows for additional structural changes during the sample period. Given that structural changes arising from shifts in the monetary policy regime are already accounted for in the baseline TVAR specification, these additional structural changes must arise from unrelated economic events.

by Hamilton (1989) for business cycle analysis. Subsequent refinements were developed by Sims, Waggoner, and Zha (2008), Sims and Zha (2006b), Rubio-Ramírez, Waggoner, and Zha (2005), and Hubrich, Waggoner, and Zha (2016). In this section we first present the general model framework and then discuss estimation, structural analysis, and applications.

18.4.1 Model Setup

Consider a reduced-form VAR(p) model

$$y_t = \nu(s_t) + A_1(s_t)y_{t-1} + \cdots + A_p(s_t)y_{t-p} + u_t, \tag{18.4.1}$$

where $\nu(s_t)$ is a $K \times 1$ intercept vector, $A_i(s_t)$ is $K \times K$ for $i = 1, \dots, p$, and $u_t|s_t \sim \mathcal{N}(0, \Sigma_u(s_t))$ is white noise with the covariance matrix depending on the state of a Markov chain s_t. The conditional distribution of u_t, given s_t, is assumed to be Gaussian. This framework encompasses a very general class of nonnormal unconditional distributions for u_t and, hence, for y_t. More details on the properties of s_t are provided later in this chapter.

This framework includes as special cases processes in which some of the model coefficients do not vary with the state of the Markov process. For example, for a conditionally homoskedastic process the covariance matrices are time invariant and $\Sigma_u(1) = \cdots = \Sigma_u(M)$. Similarly, in Chapter 14 we considered the special case of an MS-VAR model in which only the reduced-form error covariance depends on the Markov chain. Krolzig (1997) provides a full classification of these and other special cases of this framework and discusses their properties in detail.

For our purposes it is worth noting that the reduced-form setup in (18.4.1) can be transformed to the structural form

$$B_0(s_t)y_t = \nu^*(s_t) + B_1(s_t)y_{t-1} + \cdots + B_p(s_t)y_{t-p} + w_t, \tag{18.4.2}$$

where $B_i(s_t) = B_0(s_t)A_i(s_t)$ for $i = 1, \dots, p$, are the structural-form parameters. The $K \times K$ matrix $B_0(s_t)$ may be normalized to have a unit diagonal such that each equation can be written with one of the variables on the left-hand side. If $B_0(s_t)$ is unrestricted, in contrast, the error process w_t may be standardized to have unit variances. In that case, we may postulate that $w_t \sim \mathcal{N}(0, I_K)$ without loss of generality, regardless of the state of s_t. In short, w_t may be specified as invariant to the Markov chain. Alternatively, $w_t|s_t \sim \mathcal{N}(0, \Sigma_w(s_t))$ may have a diagonal covariance matrix $\Sigma_w(s_t)$ that depends on the Markov process. Since it is not uncommon in structural VAR analysis to consider just-identified structural forms where the structural- and reduced-form parameters have a one-to-one relation, we focus the discussion on a reduced-form MS-VAR analysis from which the structural-form can then be obtained by a simple transformation. This class of models does not allow for overidentifying restrictions. The latter case is discussed in detail in Sims, Waggoner, and Zha (2008) and Rubio-Ramírez, Waggoner, and Zha (2005).

Markov Chains. The model coefficients depend on the Markov chain process s_t. Generally, a process s_t that may take values $1, \dots, M$ is called a Markov chain, if the transition probabilities from one state to another depend on the most recent state only. In other words,

$$\begin{aligned} p_{ij} &\equiv \mathbb{P}(s_t = j|s_{t-1} = i) \\ &= \mathbb{P}(s_t = j|s_{t-1} = i, s_{t-2} = l, \dots), \quad i, j, l = 1, \dots, M. \end{aligned}$$

The transition probabilities are the parameters of the process and are summarized in the transition probability matrix

$$P \equiv \begin{bmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & \ddots & \vdots \\ p_{M1} & \cdots & p_{MM} \end{bmatrix}.$$

Note that the transition probability matrix is sometimes written in transposed form (e.g., Hamilton 1994, chapter 22). All elements of P are probabilities and, hence, are between zero and one, $0 \leq p_{ij} \leq 1$. Boundary values of 0 and 1 imply additional restrictions. For example, if $p_{ij} = 0$, the j^{th} state cannot be reached from state i. More generally, if P is lower block-triangular,

$$P = \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix},$$

where P_{11} contains the transition probabilities of the first M_1 states and P_{22} is the corresponding matrix for the last $M - M_1$ states, then, once one of the first M_1 states is reached, there is no possibility to return to states $m = M_1 + 1, \dots, M$. In that case the Markov chain is called reducible. Moreover, $p_{ij} = 1$ implies that state j is an absorbing state to which the process moves with probability one from state i.

Clearly, $\sum_{j=1}^{M} p_{ij} = 1$. In other words, the rows of P sum to 1 such that

$$P\mathbf{j}_M = \mathbf{j}_M,$$

where $\mathbf{j}_M$ denotes an $M \times 1$ vector of ones. Thus, P has an eigenvalue of 1 and $\mathbf{j}_M$ is the corresponding eigenvector.

In general, the transition probability matrix P contains M^2 parameters that depend on $M^2 - M$ parameters due to the restriction that $\sum_{j=1}^{M} p_{ij} = 1$. It may depend on fewer parameters, for example, if some of the transition probabilities are restricted to zero.

As mentioned earlier, MS-VAR processes are very flexible models. The description of the process in expressions (18.4.1) and (18.4.2) may seem to suggest that the process in each period has to be in one of the M states. This is not necessarily the case, however. More precisely, this DGP implies that in any given period the Markov process is in any one of the M states with a certain probability. In other words, the DGP allows for mixtures of states, which allows the process to describe far more general unconditional distributions. Even under the conditional normality assumption, for example, the unconditional distribution of y_t can have heavy tails.

18.4.2 Identification

Structural identification in MS-VAR models is discussed in detail in Rubio-Ramírez, Waggoner, and Zha (2005). Their study considers identification via short-run and long-run restrictions as well as sign restrictions. We have already seen in Chapter 14 how identification may be achieved if there is MS in the residual covariance only. Although exact identification restrictions can be stated in very general terms, in practice it is common to impose these restrictions on the coefficients of each regime separately. For example, Sims, Waggoner, and Zha (2008) consider a model for $y_t = (gdp_t, \pi_t, i_t)'$, where gdp_t is log real GDP, π_t is GDP deflator inflation, and i_t denotes the federal funds rate. They identify the shocks by imposing that $B_0(s_t)$ (and hence also $B_0(s_t)^{-1}$) are lower triangular in each state. The model is partially identified in that only the monetary policy shock is identified. The monetary policy shock is identified as an economic shock which affects the federal funds rate in the impact period, whereas it affects gdp_t and π_t only with a lag.

Rubio-Ramírez, Waggoner, and Zha (2005) also discuss the use of long-run identifying restrictions (see Chapter 10). They even provide conditions for identification if both short-run restrictions on the impact effects and long-run restrictions are considered jointly. These results, however, only pertain to imposing the long-run restrictions on the linear model for each state and do not account for the fact that the regime may change over time. Although conditioning on the economy remaining in the recession regime, for example, is common in applied work, conditioning on one regime in constructing the impulse responses is problematic because it amounts to changing the structure of the model, making the impulse response analysis subject to the Lucas Critique. Nor is it clear that such responses are relevant for understanding the response of the actual economy because agents in the real world will correctly perceive that the regime may change in the future. Once we allow for endogenous switches in the regime by computing nonlinear structural impulse responses, as discussed earlier, imposing long-run restrictions becomes impossible, because there is no longer a closed-form solution for the long-run structural responses.

Rubio-Ramírez et al. also discuss the use of sign restrictions on the impact multiplier matrix. Assuming that sign restrictions are imposed in each state, this approach poses no special difficulties. Let us assume that the covariance matrix of w_t is normalized to be an identity matrix. Then, for a given identified set of structural parameters $[B_0(m), B_1(m), \ldots, B_p(m)]$ we have to find all orthogonal matrices Q such that the impulse responses obtained from $Q[B_0(m), B_1(m), \ldots, B_p(m)]$ satisfy the desired sign restrictions for $m = 1, \ldots, M$ (see Chapter 13 for details). In contrast, the imposition of dynamic sign restrictions can only be done by simulation, once we allow for the regime to evolve over time. Either way, however, the challenge remains of how to properly evaluate the posterior distribution (or the sampling distribution) of the set of structural impulse responses.

18.4.3 Estimation

Maximum Likelihood Estimation. Using the conditional normality of u_t and w_t, the likelihood function corresponding to models (18.4.1) and (18.4.2) is

$$l = \prod_{t=1}^{T} \left(\sum_{m=1}^{M} \mathbb{P}(s_t = m | y_{t-1}, \ldots, y_1) f(y_t | s_t = m, y_{t-1}, \ldots, y_1) \right), \tag{18.4.3}$$

where

$$\begin{aligned} f(y_t|s_t &= m, y_{t-1}, \ldots, y_1) \\ &= (2\pi)^{-K/2} [\det(\Sigma_u(m))]^{-1/2} \exp\left\{ -\frac{1}{2} u_t' \Sigma_u(m)^{-1} u_t \right\} \\ &= (2\pi)^{-K/2} |\det(B_0(m))| [\det(\Sigma_w(m))]^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} w_t' \Sigma_w(m)^{-1} w_t \right\}. \end{aligned}$$

Clearly, in this case even the log-likelihood is nonlinear and the estimator must be solved for numerically. Krolzig (1997) discusses EM algorithms that simplify this optimization task for a range of important special cases. In general, ML estimation of MS-VAR models may be unreliable in small samples and may become infeasible for larger models. Moreover, inference is complicated by some of the parameters having nonstandard asymptotic distributions. This problem also undermines the reliability of frequentist model selection procedures. Moreover, little is known about the asymptotic validity of the bootstrap for MS-VAR models.

A common response in applied work to the convergence problems of the ML estimator has been the use of Bayesian estimation methods as developed in Sims, Waggoner, and Zha (2008), Sims and Zha (2006b), and Rubio-Ramírez, Waggoner, and Zha (2005). These methods are outlined in the next subsection.

Bayesian Estimation. Bayesian estimation involves setting up prior distributions for all the parameters of the model and deriving the posterior distributions of interest by simulation. For a given state of the Markov process the prior of the VAR parameters is usually set up as in the linear model. A Minnesota prior or a sum-of-coefficients prior combined with an inverse Wishart prior are common choices. Although the same prior distribution may be used in each state, this approach would be difficult to reconcile with the maintained heterogeneity of the states in the MS-VAR model.

The transition probabilities are the parameters of the Markov process. Typically, one imposes independent Dirichlet priors for the rows of P. The Dirichlet distribution is a multivariate generalization of the Beta distribution. Its density is given by

$$f(x_1,\dots,x_M;b_1,\dots,b_M)=\frac{1}{\mathrm{B}(b)}\prod_{i=1}^{M}x_i^{b_i-1}, \tag{18.4.4}$$

where $b_1,\dots,b_M>0$ are the parameters of the distribution. The density is defined for all $0\le x_1,\dots,x_M\le 1$ such that $\sum_{m=1}^{M}x_m=1$ and hence satisfies the restrictions on the row elements of the transition probability matrix P. The quantity $\mathrm{B}(b)$ is a normalizing constant that depends on the parameter vector $b=(b_1,\dots,b_M)'$. $\mathrm{B}(b)$ is a multivariate version of the Beta function that can be represented in terms of Gamma functions as

$$\mathrm{B}(b)=\frac{\prod_{i=1}^{M}\Gamma(b_i)}{\Gamma\left(\sum_{i=1}^{M}b_i\right)},$$

with $\Gamma(a)=\int_0^\infty x^{a-1}e^{-x}dx$ denoting the usual Gamma function.

If all parameters of the Dirichlet distribution are 1, i.e., $b_1=\cdots=b_M=1$, the Dirichlet distribution reduces to a multivariate uniform distribution on the unit interval [0, 1] for each of the components. If little is known a priori about the actual transition probabilities, this specification may be useful in setting up the prior. In related work, Sims, Waggoner, and Zha (2008) propose setting $b_j=1$ for the off-diagonal elements of P and using a prior parameter that reflects the persistence of the regime under consideration for the diagonal elements of P. Note that the diagonal elements of P are the probabilities of remaining in a given state. These probabilities are often rather close to 1, if the appearance of the states is not very erratic. Thus, one may want to choose the prior parameter corresponding to the i^{th} diagonal element, p_{ii}, of P such that the expected value of the marginal prior of p_{ii} is 0.9, say. Since for $b_j=1$, $j\neq i$,

$$\mathbb{E}(x_i)=\frac{b_i}{\sum_{k=1}^{K}b_k}=\frac{b_i}{b_i+(M-1)},$$

a value of 0.9 is obtained by choosing $b_i=18$ for a system with $M=3$ states.

More generally, for the i^{th} row of P the prior may be chosen as

$$f_{Pi}(p_{i1},\dots,p_{iM};1,\dots,1,b_{ii},1,\dots,1)=\frac{1}{\mathrm{B}(b)}p_{ii}^{b_{ii}-1}.$$

Thus, if the priors of the different rows are independent, the joint prior of all the rows of P is

$$\prod_{i=1}^{M} f_{Pi}(p_{i1}, \ldots, p_{iM}; 1, \ldots, 1, b_{ii}, 1, \ldots, 1).$$

Assuming independence between the elements of P and the other model parameters, the overall prior is obtained by simply multiplying this function with the prior density of the other parameters.

Combining the prior density with the likelihood yields the posterior. Unfortunately the posterior will typically not reduce to a known distribution. Sims, Waggoner, and Zha (2008) break down the posterior into conditional posteriors, at least some of which may be of a known form. For the remaining distributions they recommend the use of a Metropolis-Hastings algorithm. They do point out some problems that need to be taken into account in simulating the posterior distribution. In addition to the usual problem of choosing the sign of the shocks properly, there is also the problem of labelling the states. In other words, it is important, which state is labelled as State 1, State 2, etc. The same labelling must be used for each draw of the posterior. Sims, Waggoner, and Zha (2008) propose a complete algorithm for simulating the posterior that is designed to work even when estimating larger structural MS-VAR models. For the case of a just-identified model, draws from the reduced-form posterior can be turned into draws for the structural parameters, resulting in computational simplifications.

18.4.4 Model Selection

For MS-VAR models the number of lags and the number of Markov states have to be chosen at the specification stage. In a frequentist setting, these two quantities can be determined with penalized likelihood criteria such as the AIC subject to the caveats discussed in Chapter 2. For the MS-VAR class of models this criterion can be written as

$$\text{AIC} = -2 \log l + 2n,$$

where $\log l$ denotes the maximum of the log-likelihood function and n is the number of free parameters. The free parameters include the reduced-form MS-VAR parameters for the different regimes as well as the unconstrained elements of the transition probability matrix.[3] Similarly, other model selection criteria such as the SIC and the HQC may be used.

[3] Note that the transition probabilities add to one so that in a 2-state model we only have two unrestricted transition probabilities, in a 3-state model we have six unrestricted transition probabilities, etc.

In this context it is important to note that understating the number of Markov states may result in a change in the lag order required for the proper representation of the DGP. Krolzig (1997, chapter 3) shows that certain MS-VAR processes have a VARMA representation without Markov switching. Thus, if VAR order selection is based on a VAR model without MS, misleading results are to be expected. In other words, the choice of the number of Markov states and VAR lag-order selection cannot be seen as separate problems. Ideally a search over the full range of combinations of all plausible VAR lag orders and MS states should be performed. Unfortunately, given the computational problems in optimizing the log-likelihood function, there are limitations to the number of models that can be compared.

Psaradakis and Spagnolo (2006) report simulation results for the performance of model selection criteria for jointly choosing the number of MS states and the number of autoregressive lags for univariate AR models with MS coefficients. They conclude that penalized likelihood criteria such as the AIC are useful for determining the quantities of interest if the sample sizes are relatively large. Their study examines sample sizes of 200 and 400. A closer look at their simulation results, however, reveals that even with 400 observations and fairly simple models, the correct combination of the number of Markov states and AR order is often not detected. A reasonable conjecture is that the selection problem becomes even more difficult in multivariate models.

Of course, one could also compare models based on their relative likelihood values. Unfortunately, such tests are problematic as well because there may be unidentified parameters under the null hypothesis if a model with a smaller number of MS states is tested against a model with more states. In that situation LR tests have nonstandard distributions (for the present case, see Garcia 1998; Carrasco, Hu, and Ploberger 2014).

Finally, from a Bayesian perspective, if all models are equally likely a priori, a natural criterion for model comparison would be the marginal likelihood, as discussed in Chapter 5. The marginal likelihoods can be computed based on the algorithm discussed in the previous subsection. It should be understood, however, that the computational challenges involved make it unattractive to compare large numbers of alternative models. Thus, in practice, the choice will tend to be between a very small set of competitors.

18.4.5 Example: An MS-VAR Model of U.S. Monetary Policy

Sims, Waggoner, and Zha (2008) study a three-dimensional model of the U.S. economy comprised of the log of real GDP (gdp_t), GDP deflator inflation (π_t), and the federal funds rate (i_t). The data are quarterly and cover the period 1959q1–2005q4. The shocks are identified by assuming $B_0(s_t)$ to be a lower-triangular matrix. Sims et al. are interested in comparing models with different types of coefficient variation. Their models allow for possible variation in $B_0(s_t)$ as well as in the shock variances. The reduced-form error covariance matrix is decomposed as $B_0(s_t)^{-1}\Sigma_w(s_t)B_0'(s_t)^{-1}$, where $\Sigma_w(s_t)$ is a diagonal matrix and $B_0(s_t)$ has unit diagonal. Thus, both the impact effects of the shocks and the variances are allowed to vary across the regimes of the Markov process.

In addition, Sims, Waggoner, and Zha (2008) consider models in which the coefficient variation is restricted to subsets of the coefficients. In particular, they consider models in which only the variances vary, but the VAR slope coefficients are regime invariant, i.e., $\Sigma_w(s_t)$ varies with the Markov regimes, but $B_i(s_t) = B_i$, $i = 1, \dots, p$, and $B_0(s_t) = B_0$. They also consider models in which only the coefficients of the interest rate equation are varying, whereas the coefficients in the other equations remain regime invariant.

The priors are chosen along the lines discussed in Section 18.4.3. The specific choices of the prior parameters are not justified much. The models are compared via their marginal data density or the marginal likelihood. Sims et al. find that a model with four regimes and only the variances changing fits best under this criterion. This result suggests that the transmission of shocks has not changed, but the magnitude of the shocks has. Thus, changes in the conduct of monetary policy appear to have played a limited role during the sample period.

This result, of course, is conditional on the Markov-switching structure being an adequate representation of the data. Whether the behavior of the U.S. economy in general (and of U.S. monetary policy in particular) is as regular as implied by a Markov-switching model remains an open question, as does the sensitivity of these results to the choice of the prior. An alternative framework is the TVC-VAR model discussed in the next subsection.

Given the computational difficulties in estimating MS-VAR models, this class of models has not been widely used to date. Another example is Sims and Zha (2006b) who investigate the effects of changes in U.S. monetary policy based on a larger model including six variables. A generalization of the MS-VAR model that allows the transition probabilities to be time-varying has recently been proposed by Hubrich, Waggoner, and Zha (2016).

18.5 Time-Varying Coefficient VAR Models

There is a rapidly growing literature on time-varying coefficient VAR (TVC-VAR) models that allow for very flexible variation in the VAR coefficients, possibly including the coefficients of the error covariance matrix (e.g., Canova 1993; Canova and Gambetti 2009; Canova and Ciccarelli 2009; Cogley and Sargent 2001, 2005; Primiceri 2005; Koop and Korobilis 2009; Koop, Leon-Gonzalez, and Strachan 2009). In the next subsection, the framework developed by Primiceri (2005) is used to illustrate the problems and the potential advantages of this class of models. This framework has also served as the point of departure for many subsequent studies.

18.5.1 Model Setup

Consider the following general reduced-form process

$$y_t = \nu(t) + A_1(t)y_{t-1} + \cdots + A_p(t)y_{t-p} + u_t. \tag{18.5.1}$$

The model coefficients depend on t because they are allowed to vary over time. The error term u_t is assumed to be a zero-mean white noise process with time-varying covariance matrix, $u_t \sim (0,\ \Sigma_u(t))$. To facilitate structural analysis, the error covariance matrix is parameterized as

$$B_0(t)\Sigma_u(t)B_0'(t) = \Sigma_w(t) \quad \text{or} \quad \Sigma_u(t) = B_0(t)^{-1}\Sigma_w(t)B_0'(t)^{-1}, \tag{18.5.2}$$

where $\Sigma_w(t) = \operatorname{diag}[\sigma_1^2(t), \dots, \sigma_K^2(t)]$ is a diagonal matrix with the variances of the structural errors on the main diagonal and $B_0(t)$ is restricted such that the structural coefficients in $B_0(t)$ and $\Sigma_w(t)$ can be recovered from equation (18.5.2). For example, $B_0(t)$ may be a lower-triangular matrix with unit main diagonal,

$$B_0(t) = \begin{bmatrix} 1 & 0 & \dots & 0 \\ b_{21,0}(t) & 1 & & 0 \\ \vdots & \ddots & \ddots & \vdots \\ b_{K1,0}(t) & \dots & b_{K,K-1,0}(t) & 1 \end{bmatrix}. \tag{18.5.3}$$

Restrictions on $B_0(t)$ may be used to uniquely identify the structural shocks $w_t = B_0(t)u_t$.

The structural form of the model can be written as

$$y_t = \nu(t) + A_1(t)y_{t-1} + \cdots + A_p(t)y_{t-p} + B_0(t)^{-1}w_t.$$

Suppose that the reduced-form VAR slope coefficients are collected in the vector $\boldsymbol{\alpha}(t) = \operatorname{vec}[\nu(t), A_1(t),\ \dots, A_p(t)]$, and the unrestricted elements of $B_0(t)$ are summarized in $\mathbf{b}(t)$, where $\mathbf{b}(t) = [b_{21,0}(t),\ b_{31,0}(t),\ b_{32,0}(t),\ \dots,$ $b_{K1,0}(t), \dots, b_{K,K-1,0}(t)]'$ if $B_0(t)$ is lower triangular as in expression (18.5.3). In that case $\mathbf{b}(t)$ is the $\frac{1}{2}K(K-1)$-dimensional vector of elements below the main diagonal of $B_0(t)$. This vector is arranged row-wise such that the parameters for the individual equations are grouped together. Finally, let $\boldsymbol{\sigma}(t) = [\sigma_1(t), \dots, \sigma_K(t)]'$ be the vector of the standard deviations of the components of w_t.

The law of motion of the vectors of coefficients is assumed to be

$$\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}(t-1) + \eta_t^{\alpha}, \tag{18.5.4}$$

$$\mathbf{b}(t) = \mathbf{b}(t-1) + \eta_t^{b}, \tag{18.5.5}$$

$$\log \boldsymbol{\sigma}(t) = \log \boldsymbol{\sigma}(t-1) + \eta_t^{\sigma}, \tag{18.5.6}$$

where the error terms η_t^α, η_t^b, and η_t^σ are white noise processes. In other words, the VAR coefficients change randomly in every period, but there is considerable persistence in their movement. The first two vectors, $\boldsymbol{\alpha}(t)$ and $\mathbf{b}(t)$, follow random walks, while the process driving the standard deviations, $\boldsymbol{\sigma}(t)$, is a geometric random walk. In other words, this model allows for stochastic volatility. The error terms of the model equations are typically assumed to have a block-diagonal covariance matrix

$$\text{Cov}\begin{bmatrix} w_t \\ \eta_t^\alpha \\ \eta_t^b \\ \eta_t^\sigma \end{bmatrix} = \begin{bmatrix} \Sigma_w(t) & 0 & 0 & 0 \\ 0 & \Sigma_\alpha & 0 & 0 \\ 0 & 0 & \Sigma_b & 0 \\ 0 & 0 & 0 & \Sigma_\sigma \end{bmatrix}, \tag{18.5.7}$$

where Σ_α, Σ_b, and Σ_σ are the covariance matrices of η_t^α, η_t^b, and η_t^σ, respectively. Under the additional assumption of Gaussian errors, this structure implies independence of the individual error terms.

Defining $Z_{t-1} \equiv (1, y'_{t-1}, \dots, y'_{t-p})'$, equation (18.5.1) can be written as

$$y_t = (Z'_{t-1} \otimes I_K)\boldsymbol{\alpha}(t) + u_t. \tag{18.5.8}$$

The two equations (18.5.8) and (18.5.4) are easily recognized as a special case of a state-space model with measurement equation (18.5.8) and transition or state equation (18.5.4). Such models are well known in the time series literature. Many results exist that simplify their analysis (e.g., Lütkepohl 2005, chapter 18; Anderson and Moore 1979; Aoki 1987; Hannan and Deistler 1988; Harvey 1989; Durbin and Koopman 2012). An important tool for estimating state-space models is the Kalman filter or Kalman smoother, which refers to a very efficient algorithm for computing conditional moments such as

$$\begin{aligned} y_{t|t-1} &= \mathbb{E}(y_t | y_{t-1}, y_{t-2}, \dots, y_1) \quad \text{and} \\ \Sigma_y(t|t-1) &= \text{Cov}(y_t | y_{t-1}, y_{t-2}, \dots, y_1) \end{aligned}$$

recursively. These moments are useful for constructing estimation algorithms. We will return to the issue of estimation shortly.

The identification of structural shocks in TVC-VAR models is nontrivial in general. As Primiceri (2005) points out, there are many ways for shocks to enter the model (18.5.1). They may enter through the structural disturbances $w_t = B_0(t)u_t$ in the measurement equation or they may enter through any of the error terms in expressions (18.5.4)–(18.5.6). In the latter case it is difficult to identify the structural shocks. If w_t can be identified, conditional impulse responses can be compared over time. Obviously, in comparing the effects of structural shocks across time, the magnitude of these shocks may have to be standardized. The computation of the structural impulse responses can be based on the algorithm described in Section 18.2.2.

18.5.2 Estimation

Maximum Likelihood Estimation. If y_t is modeled as a Gaussian process, the log-likelihood function is

$$\log l = -\frac{KT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(\det(\Sigma_y(t|t-1)))$$
$$-\frac{1}{2}\sum_{t=1}^{T}(y_t - y_{t|t-1})'\Sigma_y(t|t-1)^{-1}(y_t - y_{t|t-1}). \qquad (18.5.9)$$

This likelihood function is easy to evaluate if the conditional moments $y_{t|t-1}$ and $\Sigma_y(t|t-1)$ are available. The Kalman filter provides a way of computing these expressions. Thus, for given parameter values, the numerical evaluation of the likelihood function is not difficult. Computing the maximum of the likelihood may still be a challenge, however, because the likelihood function is very nonlinear and the parameter vector can be high-dimensional. As a result, unrestricted ML estimation of TVC-VAR models tends to be problematic and typically Bayesian methods are used.

Bayesian Estimation. Primiceri (2005) describes an MCMC algorithm for the Bayesian estimation of this model. He imposes priors that are convenient for setting up such an algorithm. He also makes the simplifying assumption that Σ_b is block-diagonal with blocks corresponding to the equations of the system. In other words, for the lower-triangular $B_0(t)$ matrix in (18.5.3) the blocks have size $1 \times 1, 2 \times 2, \dots, (K-1) \times (K-1)$. Priors are required for the initial values of the coefficients and the covariance matrices of the error terms in (18.5.4)–(18.5.6). Primiceri (2005) postulates Gaussian priors for the initial conditions of $\boldsymbol{\alpha}(t)$, $\mathbf{b}(t)$, and $\log\boldsymbol{\sigma}(t)$ and inverse Wishart priors for the corresponding residual covariance matrices. Specifically, his priors are

$$\boldsymbol{\alpha}(0) \sim \mathcal{N}(\boldsymbol{\alpha}^*, V_{\boldsymbol{\alpha}}), \qquad (18.5.10)$$
$$\mathbf{b}(0) \sim \mathcal{N}(\mathbf{b}^*, V_{\mathbf{b}}), \qquad (18.5.11)$$
$$\log\boldsymbol{\sigma}(0) \sim \mathcal{N}(\log\boldsymbol{\sigma}^*, V_{\boldsymbol{\sigma}}), \qquad (18.5.12)$$
$$\Sigma_\alpha \sim \mathcal{IW}_{(pK^2+K)}(S_*^\alpha, n_\alpha), \qquad (18.5.13)$$
$$\Sigma_b \sim \mathcal{IW}_{K(K-1)/2}(S_*^b, n_b), \qquad (18.5.14)$$
$$\Sigma_\sigma \sim \mathcal{IW}_K(S_*^\sigma, n_\sigma). \qquad (18.5.15)$$

The numerical values of the prior parameters depend on the application at hand. An example is discussed in the next subsection. As noted by Primiceri (2005), a drawback of this approach is that the prior depends on the ordering of the variables. If $B_0(t)$ is lower triangular, then the elements in $\mathbf{b}(t)$ are determined by the lower triangularity of $B_0(t)$.

The details of the simulation algorithm for the joint posterior of

$$\boldsymbol{\alpha}(1),\ldots,\boldsymbol{\alpha}(T),\mathbf{b}(1),\ldots,\mathbf{b}(T),\boldsymbol{\sigma}(1),\ldots,\boldsymbol{\sigma}(T),\Sigma_\alpha,\Sigma_b,\Sigma_\sigma$$

are given in Primiceri (2005). His MCMC algorithm is based on a Gibbs sampler and proceeds in four steps:

1. Draw from the joint distribution of $\boldsymbol{\alpha}(1),\ldots,\boldsymbol{\alpha}(T)$, given $\mathbf{b}(1),\ldots,\mathbf{b}(T),\boldsymbol{\sigma}(1),\ldots,\boldsymbol{\sigma}(T)$ and values of the hyperparameters $\Sigma_\alpha,\Sigma_b,\Sigma_\sigma$. This distribution is a product of normal distributions that can be sampled with standard algorithms.
2. Draw from the joint distribution of $\mathbf{b}(1),\ldots,\mathbf{b}(T)$, given the other coefficients and hyperparameters. A similar argument as in step 1 implies that this distribution is also a product of normals.
3. Draw from the joint distribution of $\boldsymbol{\sigma}(1),\ldots,\boldsymbol{\sigma}(T)$, given $\boldsymbol{\alpha}(1),\ldots,\boldsymbol{\alpha}(T),\mathbf{b}(1),\ldots,\mathbf{b}(T)$ and the hyperparameters. For this step sampling algorithms from the stochastic volatility literature may be used (e.g, Kim, Shephard, and Chib 1998).
4. Finally, draws from the conditional posteriors of Σ_α, Σ_b, and Σ_σ may be obtained by standard methods for sampling from inverse Wishart distributions, utilizing the fact that the distributions are mutually independent.

Bayesian estimation of the TVC-VAR model considered in this subsection is also discussed in detail in Koop and Korobilis (2009) who propose a generalization of this model that allows for a slightly more general transition equation (18.5.4) (see, e.g., Canova 2007). This involves replacing expression (18.5.4) by

$$\boldsymbol{\alpha}(t)=\Pi\boldsymbol{\alpha}(t-1)+(I-\Pi)\bar{\boldsymbol{\alpha}}+\eta_t^\alpha,$$

where $\bar{\boldsymbol{\alpha}}$ is a constant vector. The additional parameter matrix Π can be treated as unknown, it can be parametrized in some way, or it simply can be treated as known. For example, one could impose restrictions that reflect the Minnesota prior. Canova (2007) discusses the case in which $\Pi = cI$ for a scalar constant c. Obviously, when $c=1$ we are back in the previously considered case with random walk dynamics for $\boldsymbol{\alpha}(t)$. When $c=0$, $\boldsymbol{\alpha}(t)$ is white noise, fluctuating around a constant mean.

Clearly, for TVC-VAR models with a large number of variables and/or a large number of autoregressive lags, the posterior becomes high-dimensional, and these simulation methods quickly become tedious or computationally infeasible. Improved Bayesian estimation methods continue to be developed (see, e.g., Koop 2013b). Nevertheless, Bayesian estimation of TVC-VAR models at this point is only feasible for small-dimensional models with a small number of autoregressive lags, restricting the use of these models in applied work. For example, it is not feasible to estimate monthly TVC-VAR models with large K and/or p.

18.5.3 Example: A TVC-VAR Model of U.S. Monetary Policy

Primiceri's (2005) study provides an example of how to apply the TVC-VAR methodology. Primiceri specifies a model with only three variables: the inflation rate (π_t), the unemployment rate (ur_t), and a short-term interest rate (r_t). The first two variables represent the non-policy sector of the economy. The interest rate equation is interpreted as a monetary policy rule. The model is semistructural. Identification is based on a recursive ordering. The model is estimated using quarterly data for 1953q1–2001q3. The first 40 observations of the sample (1953q1–1962q4) are used as a training sample to determine the numerical values for some of the parameters in the prior based on the estimate of a constant coefficient VAR model. The actual TVC-VAR analysis is conducted only on the remaining observations. The VAR order is set to $p = 2$ to keep the model tractable.

The prior is specified as follows:

$$\begin{aligned}
\boldsymbol{\alpha}(0) &\sim \mathcal{N}\left(\widehat{\boldsymbol{\alpha}}_{OLS}, 4 \cdot V(\widehat{\boldsymbol{\alpha}}_{OLS})\right), \\
\mathbf{b}(0) &\sim \mathcal{N}\left(\widehat{\mathbf{b}}_{OLS}, 4 \cdot V(\widehat{\mathbf{b}}_{OLS})\right), \\
\log \boldsymbol{\sigma}(0) &\sim \mathcal{N}\left(\log \widehat{\boldsymbol{\sigma}}_{OLS}, I_K\right), \\
\Sigma_\alpha &\sim \mathcal{IW}_{(pK^2+K)}\left(0.01^2 \cdot 40 \cdot V(\widehat{\boldsymbol{\alpha}}_{OLS}), 40\right), \\
\Sigma_\sigma &\sim \mathcal{IW}_K\left(0.01^2 \cdot 4 \cdot I_K, 4\right), \\
\Sigma_{b,11} &\sim \mathcal{IW}_1\left(0.02 \cdot V[\widehat{b}_{21,0}], 2\right), \\
\Sigma_{b,22} &\sim \mathcal{IW}_2\left(0.03 \cdot V[\widehat{b}_{31,0}, \widehat{b}_{32,0}], 3\right),
\end{aligned} \tag{18.5.16}$$

where $\widehat{\boldsymbol{\alpha}}_{OLS}$, $\widehat{\mathbf{b}}_{OLS}$, and $\widehat{\boldsymbol{\sigma}}_{OLS}$ are the LS estimates obtained from the training sample. $V(\widehat{\boldsymbol{\alpha}}_{OLS})$ is the corresponding estimator of the covariance matrix of $\widehat{\boldsymbol{\alpha}}_{OLS}$. $V[\widehat{b}_{21,0}]$ is the estimated variance of the LS estimator $\widehat{b}_{21,0}$ and $V[\widehat{b}_{31,0}, \widehat{b}_{32,0}]$ is the estimated covariance matrix of the LS estimators $(\widehat{b}_{31,0}, \widehat{b}_{32,0})$. Recall that

$$\Sigma_b = \begin{bmatrix} \Sigma_{b,11} & 0 \\ 0 & \Sigma_{b,22} \end{bmatrix}$$

is assumed to be block-diagonal. Hence, the joint distribution of Σ_b follows from that of the two diagonal blocks.

Clearly, the prior specification is arbitrary. Notably the multiplicative constants in the expressions in (18.5.16) are chosen subjectively. The choices are made for tractability and reflect the lack of prior knowledge by the analyst. As usual, the impact of the prior assumptions can be investigated by a sensitivity analysis that varies the prior parameters. Primiceri (2005) reports a range of such checks and concludes that the main results are robust to the choice of the prior.

Once the posterior simulations are available, they can be used to address a range of questions. For example, one may focus on the variability of the VAR coefficients or one may ask which model coefficients were particularly unstable over time. The evolution of $\boldsymbol{\sigma}(t)$ may also help identify periods of high volatility in shocks. As in previous subsections, the identification of the structural shocks is complicated by the absence of closed-form solutions for long-run responses. Conditional and unconditional nonlinear structural responses may be computed, as discussed in Section 18.2.2.

Primiceri (2005) concludes that the differences in how monetary policy was conducted under different Fed chairmen (Burns, Volcker, and Greenspan) "were not large enough to have any relevant effect on the dynamics of inflation and unemployment" (p. 841). Of course, that conclusion hinges on the TVC-VAR model being correctly specified and, in particular, on the specification of the monetary policy rule being appropriate for the entire estimation sample. One concern is that this model does not explicitly model other important sources of time series variation in the U.S. economy; another concern is how shocks to monetary policy can be identified from this model during time periods when the interest rate was not the policy instrument used by the Federal Reserve (see Barsky and Kilian 2001; Kozicki and Tinsley 2009); a third concern is that this semistructural model of monetary policy suffers from all the shortcomings of such models already discussed in Chapter 8.

Readers interested in other examples of the TVC-VAR methodology can find a wide range of studies that employ alternative identification strategies. For example, Benati and Mumtaz (2007) analyze a small U.S. macro model with four variables using sign restrictions and find evidence that the improved macroeconomic development after the Volcker period is due to good luck as opposed to good policy. Another example is Baumeister and Peersman (2013) who use sign restrictions to identify oil demand and oil supply shocks in a TVC-VAR analysis for the crude oil market. A common problem in such studies is how to differentiate between genuine time variation in the parameters and overfitting. Another (not unrelated) challenge is how to interpret apparent evidence of time variation from an economic point of view.

18.6 VAR Models with GARCH-in-Mean

18.6.1 Model Setup

VAR models with GARCH-in-mean have been used in a number of economic contexts. For example, Elder (2004) uses this class of models to study the effects of inflation uncertainty on output growth in the U.S. The structural form of this model is

$$B_0 y_t = \nu^* + B_1 y_{t-1} + \cdots + B_p y_{t-p} + \Lambda \boldsymbol{\sigma}_t + w_t, \qquad (18.6.1)$$

where w_t is a vector of conditionally heteroskedastic structural errors such that $w_t|\Omega_{t-1} \sim \mathcal{N}(0, \Sigma_w(t|t-1))$ with $\Sigma_w(t|t-1)$ being a conditional covariance matrix with multivariate GARCH structure and $\boldsymbol{\sigma}_t$ denoting the vector of square roots of the diagonal elements of $\Sigma_w(t|t-1)$. The term $\Lambda\boldsymbol{\sigma}_t$ is included in the conditional mean specification to capture the impact of shifts in uncertainty on the dynamics of the model observations y_t. Note that Λ may be replaced by a $K \times K$ lag polynomial $\Lambda(L)$ so that lags of conditional standard deviations of all the shocks can in principle appear in any of the VAR model equations.

Since the errors in structural model (18.6.1) are mutually uncorrelated, the multivariate GARCH structure reduces to a system of individual GARCH processes for each of the error terms. This fact allows us to impose a diagonal GARCH structure such that $\Sigma_w(t|t-1)$ is a diagonal matrix with diagonal elements

$$\sigma_{kt}^2 = c_k + \gamma_k w_{k,t-1}^2 + g_k \sigma_{k,t-1}^2, \quad k = 1, \dots, K. \tag{18.6.2}$$

Of course, higher-order GARCH models could also be considered, although this is rarely done in practice.

Whereas conditional heteroskedasticity in the form of GARCH errors poses no fundamental problems in otherwise linear VAR models, the presence of the conditional standard deviation in the conditional mean specification renders the conditional mean nonlinear and conventional structural impulse response analysis invalid. Conditional and unconditional nonlinear structural impulse responses may be computed by simulation, as discussed in Section 18.2.2. In related work, Elder (2003) proposes a closed-form solution of the conditional nonlinear structural response for the VAR-GARCH-in-mean model that allows structural shocks to affect the conditional expectation of y_{t+h} through the conditional variance (see also Elder 2004; Elder and Serletis 2010).

18.6.2 Estimation

VAR models with GARCH-in-mean are typically estimated by Gaussian ML methods. Apart from a constant term, the log-likelihood function is

$$\begin{aligned}\log l = &-T \log|\det(B_0)| \\ &-\frac{1}{2}\sum_{t=1}^{T}\left[\log(\det \Sigma_w(t|t-1)) + w_t'\Sigma_w(t|t-1)^{-1}w_t\right].\end{aligned} \tag{18.6.3}$$

Because $w_t = B_0 y_t - \nu^* - B_1 y_{t-1} - \cdots - B_p y_{t-p} - \Lambda\boldsymbol{\sigma}_t$, the log likelihood is a highly nonlinear function of the parameters. As a result, ML estimation may be problematic for models of larger dimension, models with larger orders, and/or models with more complicated forms of conditional heteroskedasticity. VAR Models with GARCH-in-mean are usually assumed to satisfy the conditions required for ML estimators to have standard asymptotic properties. In addition to the usual identification restrictions for specifying the structural shocks, w_t, this means that we also need to ensure the identification of the GARCH parameters which in turn may require further identification conditions if more complicated GARCH structures are allowed for (see, e.g., Lütkepohl 2005, chapter 16).

18.6.3 Example: The Effect of Oil Price Uncertainty on U.S. Real Output

This class of models is best illustrated in the context of a study by Elder and Serletis (2010) that investigated the impact of oil price volatility on U.S. output growth based on a VAR model with GARCH-in-mean. This study was a major improvement on earlier studies that incorrectly treated changes in the price of oil as strictly exogenous with respect to the U.S. economy.

Elder and Serletis' baseline model was bivariate and included real output growth and the percent change in the real price of oil. They used quarterly data for 1974q2–2008q1, fit a model with diagonal GARCH specification, and identified the shocks by letting B_0 be lower triangular. In other words, they treated the price of oil as predetermined with respect to U.S. real output growth, consistent with evidence in Kilian and Vega (2011). They found that increased oil price volatility has a negative effect on U.S. real output, as measured by the coefficient of the conditional standard deviation in the real output equation of the VAR model. Negative effects were also found for various other measures of output.

This evidence is economically important because it implies that output contracts sharply in response to positive oil price shocks, but fails to expand in response to negative oil price shocks. This asymmetry result is surprising because, as we will see later in this chapter, other less parametric studies looking for asymmetric responses to positive and negative oil price shocks have failed to find such evidence.

Elder and Serletis interpret their evidence as being supportive of the real options theory of Bernanke (1983), which their model was designed to capture. Real options theory states that firms will delay investment in response to increased uncertainty about the price of crude oil to the extent that the cash flow of investment projects depends on the price of crude oil. Reductions in investment in turn have a recessionary effect.

Elder and Serletis' interpretation has been challenged by a number of studies (see Kilian 2014). One concern has been that what matters for real options theory is the uncertainty about the price of oil at the horizons that matter for investment decisions. Most investment projects take at least one year to become operational. For a typical investment project (such as an auto maker deciding whether to build a new plant producing sports utility vehicles) the relevant horizon for the cash flow will typically extend from one year after the investment project is approved to several years after the project has started operating.

For concreteness, suppose that the decision to invest is based on the cash flow within the first five years of the plant becoming operational. Given the fact that the conditional variance from monthly or quarterly GARCH models quickly converges to the time-invariant unconditional variance at longer prediction horizons, one would not expect the conditional variance between years 1 and 6 to change very much at all. Thus, there is little reason for such an investment project to be delayed in response to an increase in the GARCH estimate of the monthly conditional variance.

Moreover, the share of investments whose cash flow depends heavily on the price of crude oil is quite small outside the oil industry, making it even less likely that the effects on aggregate U.S. real output would be large. For these two reasons, there is every reason to believe that whatever effect Elder and Serletis estimated is unrelated to real options theory and may reflect model misspecification.

A second concern is that GARCH models are inherently backward-looking, which makes them potentially poor measures of uncertainty. For example, GARCH models imply a sharp increase in uncertainty after the sharp fall in the price of oil in 1986. This fall in the price of oil was caused by Saudi Arabia's decision to no longer restrict its oil production, reversing an earlier policy decision to curb Saudi oil production in an effort to prop up the price of crude oil (see Alquist, Kilian, and Vigfusson 2013). Most observers would have interpreted this policy reversal as a reduction in uncertainty rather than an increase in uncertainty, because this event revealed Saudi Arabia's and hence OPEC's inability to control the price of oil, which in turn alleviated fears about what OPEC might do in the future to raise the price of oil.

A third and related concern is that the GARCH-in-mean specification imposes the assumption that the conditional mean and the conditional variance of y_t are driven by one and the same structural shock. As observed by Jo (2014), a more credible specification would be a VAR model with stochastic volatility that allows the conditional variance to evolve independently from the conditional mean. This modification, for example, would allow oil price uncertainty to fall, as the price of oil dropped in 1986.

The challenge in implementing this alternative modeling approach in practice is to find an independent measure of oil price uncertainty. Based on a novel measure of short-term oil price uncertainty, Jo finds that the recessionary effects of increases in uncertainty are considerably smaller than reported by Elder and Serletis. Even her evidence does not decisively resolve this question, however. The fundamental challenge for the stochastic volatility approach is that the economically relevant measure of uncertainty remains medium-term uncertainty (defined as uncertainty at horizons between one year and five years, for example), but measures of medium-term stochastic volatility are not likely to be exogenous, complicating the identification of the responses to uncertainty shifts.

A final concern is that the price of oil responds to actual and expected fluctuations in global real economic activity. As a result, oil price uncertainty also tends to reflect macroeconomic uncertainty which directly affects the cash flow of many investment projects. This fact may help explain the unexpectedly large effects of oil price uncertainty reported in this literature.

18.7 Other Nonlinear Models

In this section we review several alternative nonlinear models that may become useful for structural analysis in the future, although their use has been limited so far. We focus on two classes of models. The nonparametric models considered in Section 18.7.1 allow flexible approximations to very general nonlinear functions. They are designed for situations in which the specific nonlinear form of the model is unknown. Estimates of nonparametric models may also provide empirical support for the use of a specific nonlinear functional form in structural modeling. In contrast, the noncausal VAR models discussed in Section 18.7.2 are intended for situations in which economic agents have more information than the econometrician. If the predictions of the economic agents differ from those implied by the econometric model, this fact has to be taken into account in setting up the model (see also Chapter 17).

18.7.1 Nonparametric VAR Analysis

Prediction involves capturing the conditional means, conditional variances, or the conditional densities of a vector of time series, given its past realizations. These characteristics are also of interest in reduced-form modeling. If little is known about the DGP, it makes sense to explore very general potentially nonlinear functional forms. This situation calls for the use of nonparametric estimation methods. For example, a nonparametric analysis of the conditional mean would start from the model

$$y_t = \mu(y_{t-1}, y_{t-2}, \dots) + u_t, \tag{18.7.1}$$

where u_t is a martingale difference series with respect to $\{y_{t-1}, y_{t-2}, \dots\}$. In that case, $\mu(\cdot)$ represents the conditional expectation in period t, given past observations $y_{t-1}, y_{t-2}, \dots$. The conditional expectation is the minimum mean squared error (MSE) one-step ahead predictor for y_t. For a finite-order VAR(p) process

$$\mu(y_{t-1}, y_{t-2}, \dots) = \nu + A_1 y_{t-1} + \cdots + A_p y_{t-p}.$$

In parametric time series analysis, the function $\mu(\cdot)$ is chosen from some parametric class. A specific model within that class is then obtained by specifying the finite-dimensional set of model parameters associated with that function. In contrast, nonparametric time series analysis allows $\mu(\cdot)$ to be from a more flexible class of functions. We approximate $\mu(\cdot)$ such that the approximation error declines with the sample size.

There are several techniques that can be used for this purpose. For example, we may construct local approximations of $\mu(\cdot)$ in the neighborhood of any given argument by decreasing the neighborhood (and thereby improving the approximation), as the sample size increases. In this approach the number of lagged values of y_t is usually fixed. In other words, $\mu(y_{t-1}, y_{t-2}, \dots)$ is replaced by $\mu(y_{t-1}, \dots, y_{t-p})$ for given p. Alternatively, we may construct global approximations that involve parametric functions $\mu_T(\cdot)$. The number of parameters in that function (and hence the flexibility of the function) increases with the sample size T. The sequence of functions $\mu_T(\cdot)$ is chosen such that it approaches $\mu(\cdot)$ asymptotically, given the metric chosen by the researcher. This fact allows one to increase the number of lagged y_t with the sample size T without having to assume the existence of a particular finite-order parametric VAR model.

If the conditional volatility is of interest in addition to the conditional mean, the framework in (18.7.1) must be replaced by

$$y_t = \mu(y_{t-1}, y_{t-2}, \dots) + \Sigma^{1/2}(y_{t-1}, y_{t-2}, \dots)w_t \tag{18.7.2}$$

where $\Sigma^{1/2}(\cdot)$ is defined by $\Sigma(\cdot) = \Sigma^{1/2}(\cdot)\Sigma^{1/2}(\cdot)$, $\Sigma(\cdot)$ denotes the conditional covariance matrix of the process in period t, given the information from previous periods, and w_t is a white noise process with mean zero and identity covariance matrix I_K. As for the conditional mean model, several nonparametric approaches exist for jointly estimating $\mu(\cdot)$ and $\Sigma^{1/2}(\cdot)$. Of course, it is also possible to specify a parametric form of one of the two objects of interest and to model the other one nonparametrically.

More generally the complete predictive (conditional) density $f(y_t|y_{t-1}, y_{t-2}, \dots)$, may be of interest, in particular, if higher order moments are relevant for the analysis. For this case as well a number of nonparametric approaches have been proposed. We first deal with local smoothing methods, before presenting the so-called semi-nonparametric (SNP) approach as one possible example of a global approximation method. Our analysis merely conveys the general ideas. A more detailed review of nonparametric modeling techniques for univariate time series with additional references can be found in Härdle, Lütkepohl, and Chen (1997). Many of these techniques generalize to the multivariate case, but their applicability may be limited by the curse of dimensionality.

Local Methods.

Estimator of a joint density. Suppose we are interested in the joint density of the Kp-dimensional vector $Y_t = (y_t', \dots, y_{t-p+1}')'$,

$$f(Y_t) = f(y_t, \dots, y_{t-p+1}).$$

Given a sample of time series variables $y_1, \dots, y_T$ plus all required presample values, one way of estimating $f(\cdot)$ is the kernel estimator

$$\hat{f}_T(Y) = \frac{1}{Th^{Kp}} \sum_{t=1}^{T} \mathbf{K}\left(\frac{Y - Y_t}{h}\right), \tag{18.7.3}$$

where Y is a Kp-dimensional vector, $\mathbf{K}(\cdot)$ is a multivariate kernel function, and the scalar $h > 0$ in this subsection is used to denote the so-called bandwidth. For example, $\mathbf{K}(\cdot)$ may be a multivariate standard normal density or, more generally, it may be expressed as a product

$$\mathbf{K}(\cdot) = \prod_{j=1}^{Kp} \mathbf{K}_j(\cdot),$$

where the $\mathbf{K}_j(\cdot)$ are univariate kernel functions. The bandwidth h is chosen as a decreasing function of the sample size T such that the neighborhood from which we estimate the functional value shrinks, as the sample size increases. Clearly, for the estimate to become increasingly accurate with increasing sample size, we require that more and more observations lie in the neighborhood of the function value under consideration. Often $\mathbf{K}(\cdot)$ will have compact support. Under suitable mixing conditions, and appropriate kernel functions and bandwidth choice, the estimator $\hat{f}_T(Y)$ is consistent. For this result and related asymptotic properties see Robinson (1983). For an expository discussion of the univariate case see Tschernig (2004).

Once we have obtained the joint density of the Y_t, the conditional densities and their conditional moments can be derived.

Conditional moments. The conditional mean

$$\mu(y_{t-1}, \dots, y_{t-p}) = \mathbb{E}(y_t | y_{t-1}, \dots, y_{t-p}) = \int y f(y | y_{t-1}, \dots, y_{t-p}) dy$$

may serve as the one-step ahead predictor in period $t - 1$. For a given $Kp \times 1$ vector Y, the conditional expectation function may be estimated as

$$\widehat{\mu}(Y_{t-1}) = \frac{\sum_{t=1}^{T} \mathbf{K}\{(Y - Y_{t-1})/h\} y_t}{\sum_{t=1}^{T} \mathbf{K}\{(Y - Y_{t-1})/h\}}, \tag{18.7.4}$$

where $\mathbf{K}(\cdot)$ is again a kernel function and h is the bandwidth.

Predictive density. Similarly, the conditional density

$$f(y_t|Y_{t-1}) = f(y_t, y_{t-1}, \dots, y_{t-p})/f(y_{t-1}, \dots, y_{t-p})$$

may be considered. This predictive density may be estimated analogously to the joint density as

$$\hat{f}(y_t|Y_{t-1}) = \frac{h^{-K}\sum_{t=1}^{T}\mathbf{K}\{(Y^* - Y_{t-1}^*)/h\}}{\sum_{t=1}^{T}\mathbf{K}\{(Y - Y_{t-1})/h\}} = \frac{\hat{f}_T(Y^*)}{\hat{f}_T(Y)}, \qquad (18.7.5)$$

where Y^* and $Y_t^* = (y_t', \dots, y_{t-p}')'$ are $K(p+1)$-dimensional vectors.

Instead of a kernel estimator, local polynomial approximations could also be used, as proposed by Härdle and Tsybakov (1997) and Härdle, Tsybakov, and Yang (1998), among others. All these approaches require larger samples than may be available in applied work. In VAR modeling, some simplifications therefore may be helpful, as discussed next.

An additive nonparametric VAR model. Because reliably estimating fully flexible functions by nonparametric local methods requires large samples, and because such samples are often not available in macroeconomic analysis, Jeliazkov (2013) proposes a more parsimonious nonlinear VAR model with additive errors of the form

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Kt} \end{pmatrix} = \begin{pmatrix} a_{11,1}(y_{1,t-1}) + \cdots + a_{1K,1}(y_{K,t-1}) \\ \vdots \\ a_{K1,1}(y_{1,t-1}) + \cdots + a_{KK,1}(y_{K,t-1}) \end{pmatrix} + \cdots$$
$$+ \begin{pmatrix} a_{11,p}(y_{1,t-p}) + \cdots + a_{1K,p}(y_{K,t-p}) \\ \vdots \\ a_{K1,p}(y_{1,t-p}) + \cdots + a_{KK,p}(y_{K,t-p}) \end{pmatrix} + u_t, \qquad (18.7.6)$$

where the error process u_t is a white noise process and the $a_{ij,k}(\cdot)$ are fully flexible functions of the lagged variables. The advantage of this setup is that all these functions are only one-dimensional. Jeliazkov (2013) proposes a Bayesian approach to estimating these functions nonparametrically and for choosing a suitable lag order. He also proposes a number of extensions of the model. For example, changes in volatility can be accommodated. His algorithm relies on smoothness priors that have been used also by a number of other authors (see Jeliazkov (2013) for further references).

As an example consider a quarterly model of the U.S. economy. Let $y_t = (\Delta gdp_t, ur_t, r_t, \pi_t)'$, where Δgdp_t is real GDP growth, ur_t is the unemployment rate, π_t is consumer price inflation, and r_t is the three-month Treasury bill rate. Jeliazkov fits a nonparametric VAR(1) model to data for 1948q1–2005q1. He shows that a number of the functions $a_{ij,k}(\cdot)$ are not very different from linear functions. Such results may be used to simplify the model. He also finds, however, evidence of nonlinearities that must be accounted for to avoid misspecification bias. Notably some of the functions involving lags of inflation, $a_{i4,1}(\pi_{t-1})$, appear to be nonlinear in the Δgdp_t and ur_t equations.

Analysis along the lines of this example may in some cases suggest a more tightly parameterized nonlinear specification that can be used for structural analysis, as discussed in earlier sections of this chapter. It may also be used to assess whether a given parametric nonlinear specification is consistent with the data.

Global Approximation – the SNP Approach. In related work, Gallant and Tauchen (1989) use Hermite expansions to approximate the one-step ahead conditional density of the DGP of a multiple time series variable, y_t, given its past. This approach is based on the fact that a large class of density functions, $f(z)$, is proportional to $[\psi(z)]^2\phi(z)$, where $z = \Sigma_y^{-1/2}(y - \mu_y)$, with μ_y and $\Sigma_y^{-1/2}$ denoting the location and scale parameters, respectively, of the distribution, with $\psi(z)$ denoting a multivariate polynomial of possibly infinite degree r, and with $\phi(z)$ representing the multivariate standard normal density. Dividing $[\psi(z)]^2\phi(z)$ by a normalizing constant, this expression is just the Hermite expansion of $f(z)$. Hence, the density may be written as the product of a standard normal density and the square of a polynomial.

Suppose we are interested in the conditional density $f(y_t|\, y_{t-1}, y_{t-2}, \dots)$. Then

$$f(y_t|y_{t-1}, y_{t-2}, \dots) \propto [\psi(z_t)]^2\phi(z_t), \tag{18.7.7}$$

where $z_t = \Sigma_t^{-1/2}(y_t - \mu_t)$ with μ_t and $\Sigma_t^{-1/2}$ being location and scale parameters, respectively, of the conditional distribution. The former is assumed to be a linear function of the past, $\mu_t = \nu + A_1 y_{t-1} + \dots + A_p y_{t-p}$, while the latter may be modeled as an ARCH-type function (see Gallant, Hsieh, and Tauchen 1991; Gallant and Tauchen 1994). Regardless of the specification, the location and scale parameters μ_t and $\Sigma_t^{1/2}$ are modeled parametrically, whereas higher-order moments are captured by the polynomial. Letting the polynomial degree increase with the sample size makes this approach nonparametric. The approach is also known as semi-nonparametric (SNP) in the literature because it combines parametric with nonparametric elements.

To achieve a flexible adjustment of the model to higher-order dynamics the coefficients of the polynomial $\psi(\cdot)$ may be made dependent on lagged y_t. Of course, for polynomial degree $r = 0$ we have

$$f(y_t|y_{t-1}, y_{t-2}, \dots) \propto \phi\left(\Sigma_t^{-1/2}(y_t - \nu - A_1 y_{t-1} - \dots - A_p y_{t-p})\right)$$

so that a linear VAR(p) process with conditionally heteroskedastic error term emerges as a special case of the SNP model.

For given values of the integer parameters specifying the lag lengths and the polynomial degrees, this model can be estimated by maximizing the normalized log-likelihood,

$$\log l(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|y_{t-1}, \dots, y_{t-p}; \theta).$$

Asymptotic properties of this estimation procedure are provided in Gallant and Nychka (1987) who allow the order of the Hermite expansion to increase with the sample size. Gallant and Tauchen (1994) propose model selection criteria for choosing the integer parameters of the model.

Ideally the SNP model reduces to a model that can be handled within a standard linear or nonlinear parametric VAR framework. Even if a more general model is required for an adequate description of the DGP, it may still be possible to derive the structural form of such a model. At this point, however, there are few empirical examples of such studies.

18.7.2 Noncausal VAR Models

An important characteristic of stationary linear VAR processes is that the variables y_t depend on lagged values $y_{t-1}, y_{t-2}, \dots$, plus a white noise error term such that

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

If the autoregressive operator $A(z) = I_K - A_1 z - \cdots - A_p z^p$ has no roots in and on the complex unit disk, i.e.,

$$\det(A(z)) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1,$$

the process has a one-sided MA representation,

$$y_t = A(L)^{-1} u_t = \sum_{i=0}^{\infty} \Phi_i u_{t-i},$$

that depends only on past and present error terms u_{t-i}, $i = 0, 1, \dots$. Such processes are called causal because y_t is determined by its past.[4] In this case, if u_t is a martingale difference process, the conditional expectation

$$\mathbb{E}(y_t|y_{t-1}, y_{t-2}, \dots)$$

is the best prediction of y_t in period $t-1$ and u_t is the one-step ahead prediction error,

$$u_t = y_t - \mathbb{E}(y_t|y_{t-1}, y_{t-2}, \dots).$$

[4] This terminology is not to be confused with the notion of causality discussed in Chapter 7.

As discussed in Chapter 17, in some economic models current values of y_t also depend on future values of y_t (see Hansen and Sargent 1980, 1991). Such a situation arises when economic agents have more information than the econometrician. In that case, predictions based on the econometrician's VAR model will not be as accurate as the expectations of the economic agents who anticipate future values of some of the model variables and take this information into account in their decision-making. In this situation, the class of causal VAR models is too restrictive. Future values of y_t have to be allowed for in the DGP, resulting in a noncausal model.

Noncausal univariate AR models have been considered by Brockwell and Davis (1987, chapter 3), Breidt, Davis, Lii, and Rosenblatt (1991), and Lanne and Saikkonen (2011), for example. There are several proposals for the construction of multivariate noncausal models. For example, a noncausal VAR model suggested by Lanne and Saikkonen (2013) takes the form

$$A(L)C(L^{-1})y_t = u_t, \tag{18.7.8}$$

where $A(L) \equiv I_K - A_1L - \cdots - A_pL^p$ is a matrix polynomial in the lag operator L and $C(L^{-1}) \equiv I_K - C_1L^{-1} - \cdots - C_qL^{-q}$ is a matrix polynomial in the inverse lag operator L^{-1} that is defined such that $L^{-1}y_t = y_{t+1}$. For $p = q = 1$, $A(L)C(L^{-1}) = I_K + A_1C_1 - A_1L - C_1L^{-1}$, and the process (18.7.8) simplifies to

$$(I_K + A_1C_1)y_t = A_1y_{t-1} + C_1y_{t+1} + u_t.$$

Thus, y_t depends on lagged and future values. Clearly, if $q = 0$ and, hence, $C(L^{-1}) = I_K$, y_t reduces to a standard causal VAR process.

Lanne and Saikkonen (2013) assume that both operators $A(L)$ and $C(L)$ are invertible with all roots outside the complex unit circle such that

$$\det(A(z)) \neq 0 \quad \text{and} \quad \det(C(z)) \neq 0 \quad \forall z \in \mathbb{C}, |z| \leq 1.$$

Hence, $x_t = C(L^{-1})y_t$ is a stable, stationary process,

$$A(L)x_t = u_t$$

with MA representation

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

The process y_t is also stationary in this case. However, in general y_t is easily seen to have a two-sided MA representation in terms of the u_t errors,

$$y_t = \sum_{i=-\infty}^{\infty} \Psi_i u_{t-i}.$$

Moreover, unlike in the standard causal VAR case,

$$y_t - \mathbb{E}(y_t|y_{t-1}, y_{t-2}, \dots) \neq u_t.$$

Hence, the VAR model errors u_t are not one-step ahead prediction errors in general. In other words, u_t is y_t-nonfundamental in the terminology used in Chapter 17. In fact, there is in general no closed-form expression for the conditional expectation $\mathbb{E}(y_t|y_{t-1}, y_{t-2}, \dots)$. This conditional expectation is a highly nonlinear function of lagged y_t and, hence, predictions may have to be computed by simulation. This is also the reason why noncausal models are discussed in this chapter on nonlinear VAR models. Since impulse responses can also be viewed as conditional expectations based on past information, they are nonlinear functions of the shocks as well.

Lanne and Saikkonen (2013) develop an ML procedure for estimating the parameters of their model. They also provide advice on how to specify noncausal VAR models. An important point to note in this context is that stationarity of y_t implies that there exists a one-sided Wold MA representation that generates the same autocovariance structure as that of y_t. Thus, if y_t is Gaussian and, hence, also u_t is Gaussian, the DGP is indistinguishable from a causal process with the same first and second moments. Since Gaussian processes are fully determined by the first and second moments, noncausal VAR models are of interest only if the distribution of y_t is nonnormal.

As mentioned earlier, there are other ways of specifying noncausal VAR models. So far we have focused on the framework of Lanne and Saikkonen (2013) who use the multiplicative VAR operator in (18.7.8). Alternatively, one may consider a general additive VAR operator with leads and lags and write $D(L)y_t = u_t$ with

$$D(L) = I_K - D_1 L - \cdots - D_p L^p - D_{-1} L^{-1} - \cdots - D_{-r} L^{-r}.$$

Yet another approach is to consider the process

$$A^*(L)y_t = u_t \tag{18.7.9}$$

with one-sided operator $A^*(L) = I_K - A_1^* L - \cdots - A_p^* L^p$. This process is allowed to have roots inside (but not on) the unit circle such that

$$\det(A^*(z)) \neq 0 \quad \forall z \in \mathbb{C}, |z| = 1,$$

but there may be some z_0 with modulus less than 1 such that $\det(A^*(z_0)) = 0$. In other words, only roots on the unit circle are excluded. If there are roots inside the unit disk, then y_t is noncausal. This framework has been used by Davis and Song (2012), for example. Their framework encompasses processes that are not covered by the multiplicative formulation used by Lanne and Saikkonen (2013). The reverse, however, is also true in that there are multiplicative models that cannot be embedded in (18.7.9), so it may matter which framework is used.

Applications of this class of models can be found in Lanne and Saikkonen (2013) and Davis and Song (2012). These studies consider term structure models of interest rates with two factors, one defined as the difference between a long-term and a short-term interest rate and the other defined as the change in the short-term interest rate. Such systems have been used in the past to test the expectations hypothesis of the term structure within the framework of causal models. Lanne and Saikkonen and Davis and Song both find evidence for a noncausal model. Their result casts doubt on previous studies of the expectations hypothesis of the term structure based on causal VAR models.

18.8 Discussion of Nonlinear VAR Modeling

This chapter has shown that there are many types of nonlinear VAR models, raising the question of how to choose among these specifications. It is generally good advice to use nonlinear models only when there is a compelling economic rationale for nonlinearities. Often the economic context dictates the type of nonlinearity that is called for. For example, institutional constraints may suggest the existence of thresholds, a recurring pattern of expansions and contractions may suggest a Markov-switching model, or theoretical economic models may suggest a GARCH-in-mean VAR specification. Knowing what type of nonlinearity one is interested in also makes it easier to determine whether there is enough variation in the sample to accurately estimate the nonlinear phenomenon of interest. One would not expect to be able to estimate accurately any type of nonlinearity that occurs only rarely in the data or that occurs for the first time at the very end of the sample such as the zero lower bound on U.S. interest rates after 2007.

There are situations, however, when the choice of the nonlinear specification is not as clear-cut, because there may be more than one economic rationale to consider. A case in point is the literature on asymmetric responses to positive and negative oil price shocks. Such asymmetries may be motivated based on any number of mechanisms including real options theory, precautionary savings models, reallocation models with economic frictions, or simply behavioral models (see Kilian 2014). In such a situation it can be difficult to separate the empirical content of alternative economic models, and any nonlinear model focusing on one economic mechanism only, while excluding the other explanations, is likely to be misspecified. For example, a VAR model with GARCH-in-mean errors would be inherently misspecified in the presence of other sources of asymmetric responses. Moreover, the DGP may be subject to additional nonlinearities not considered by standard economic models, which also would undermine the validity of the GARCH-in-mean VAR model. Examples are changes in the share of oil in real output or the development of alternative sources and uses of energy.

In this case, a less parametric approach that encompasses many competing economic explanations such as the TVC-VAR model may be preferable, provided that approach is computationally feasible. The approach of fitting TVC-VAR models is not without risk, however, because nonlinear models tend to overfit in small samples, especially in the presence of large economic events. It is, of course, possible to guard against overfitting by tightening the priors used in estimating TVC-VAR models, but the use of such prior information becomes problematic if the prior is chosen for convenience rather than on economic grounds, as is typically the case in practice. Nor is the standard TVC-VAR model with slope parameters evolving according to unrestricted random walks designed to approximate stationary nonlinear processes. Its use only makes sense if there is some credible source of nonstationarity. Another caveat is that the TVC-VAR model is not designed for situations where the model is time-invariant for part of the sample and subject to structural change for the remainder of the sample.

More generally, it is important to keep in mind that the greater flexibility of nonlinear models in fitting the data need not translate into meaningful estimates. Of course, the reverse is also true in that nonlinearities may not be apparent if there is not enough variation in the data. Only large economic events such as a major recession or a stock market crash may reveal nonlinear behavior in the data. Thus, to the extent possible, estimates of such models should be checked against extraneous evidence. Users of these models need to make the case that the timing, direction, and magnitude of the variation in the structural model parameters is economically plausible.

It is also worth pointing out that the distinction between a nonlinearity and a structural break may not be easy to make on empirical grounds. Some of the models presented in this chapter can in fact be viewed as models for structural change. For example, TVC-VAR models designed to capture smooth structural change in practice often are able to detect discrete structural changes as well. Likewise, an MS-VAR model may be suitable for detecting structural breaks. The MS-VAR model assumes that the economy can be in a finite number of alternative states. It assigns, on the basis of the sample information, probabilities to these states in each period of the sample. Although typically the system will alternate between states, such a model can also capture a single structural break.

Whether such a break should be modeled as a deterministic structural break or not, is mainly a question of its economic interpretation. Deterministic breaks tend to be associated with discrete changes in institutions that are effectively permanent such as a currency union or the liberalization of a market. Most structural changes that economists are concerned with are likely to be stochastic rather than deterministic and continuous (gradual and smooth) rather than discrete (sudden and abrupt). The key difference is that deterministic breaks are assumed not to occur again in the future, whereas stochastic breaks would be expected to occur according to some possibly exogenous stochastic process. This fact makes a difference for the construction of structural impulse responses as well as for forecasting.

Finally, it should be noted that many events commonly associated with structural breaks in the literature such as World War II or the Great Depression are not likely sources of permanent structural changes. Standard economic theory, for example, implies that a temporary surge in government spending triggered by a war would have a temporary effect on the level of real output. Likewise, to the extent that the Great Depression was caused by monetary policy, it should represent a temporary contraction rather than a permanent structural shift. Such large transitory dynamics may be difficult to distinguish from nonlinearities or structural change even in moderately large samples (see Kilian and Ohanian 2002).

18.9 Linear Structural Models with Nonlinear Transformations of the Variables

Generalizing VAR models to incorporate nonlinearities does not necessarily imply specifications that are nonlinear in the model parameters. Instead models may be linear in the parameters, but nonlinear in the variables. This point is best illustrated by the literature on modeling the transmission of oil price shocks. It is common in applied work to rely on nonlinear transformations of the price of oil in modeling the relationship between the price of oil and the domestic economy. The objective of such transformations is to capture asymmetries and other potential nonlinearities in the transmission of oil price shocks to the economy. Often the transformations used involve some censoring of the data. For example, Mork (1989) considered the percent increase in the real price of oil, defined as $\Delta r_t^+ = \max(0, \Delta r_t)$, where Δr_t is the percent change in the real price of oil and r_t is the log level of the real price of oil. Hamilton (1996) proposed replacing the Mork measure by the 1-year net oil price increase measure, defined as $\Delta r_t^{net,+,1yr} = \max(0, r_t - r_t^*)$, where r_t^* denotes the highest log real price of oil over the preceding year. Finally, Hamilton (2003) recommended the 3-year net oil price increase measure based on the maximum price of oil over the preceding three years. The latter specification is the oil price transformation used most commonly in applied work.

Our discussion, without loss of generality, focuses on the 3-year net oil price increase measure. Both nominal and real oil price transformations have been used in applied work, although only the latter specification is consistent with economic theory. We follow the recent literature in focusing on the real price of crude oil and its relationship with U.S. real GDP.

18.9.1 The Censored Oil Price VAR Model

Traditionally, researchers interested in the relation between the real price of oil and U.S. economic growth have relied on reduced-form VAR models of the form

$$\begin{aligned}\Delta r_t^{net,+,3yr} &= \nu_1 + \sum_{i=1}^{p} a_{11,i}\Delta r_{t-i}^{net,+,3yr} + \sum_{i=1}^{p} a_{12,i}\Delta gdp_{t-i} + u_{1t},\\ \Delta gdp_t &= \nu_2 + \sum_{i=1}^{p} a_{21,i}\Delta r_{t-i}^{net,+,3yr} + \sum_{i=1}^{p} a_{22,i}\Delta gdp_{t-i} + u_{2t},\end{aligned} \tag{18.9.1}$$

where Δgdp_t refers to the growth rate of U.S. real GDP and $\Delta r_t^{net,+,3yr} = \max(0, r_t - r_t^*)$, where r_t^* denotes the highest log real price of oil over the preceding three years, as defined earlier. Identification is achieved by imposing that the real price of oil is predetermined with respect to U.S. real GDP (see Kilian and Vega 2011). In practice, the structural model is inferred by applying a lower-triangular Cholesky decomposition to the reduced-form white noise covariance matrix Σ_u under the assumption that the real price of oil is predetermined. Impulse responses are constructed exactly as in linear structural VAR models. Examples of this approach include Hamilton (1996), Bernanke, Gertler, and Watson (1997), and Ramey and Vine (2011), among many others.

These models became popular among macroeconomists because of their ability to generate "better looking impulse responses", referring to the model's ability to generate larger economic declines in response to positive oil price shocks than linear VAR models (see Bernanke, Gertler, and Watson 1997). This feature is an artifact of the model specification, however, rather than driven by the data.

Kilian and Vigfusson (2011a) demonstrate that asymmetric models of the transmission of oil price shocks cannot be represented as censored oil price VAR models of the form (18.9.1). Whether the DGP is symmetric or asymmetric in oil price increases and oil price decreases, the censored oil price model is inherently misspecified. This misspecification renders the parameter estimates inconsistent and inference invalid. Moreover, standard approaches to generating structural impulse responses from model (18.9.1) are invalid because of the presence of censored variables. As a result, the censored oil price VAR model tends to substantially overstate the recessionary impact of positive oil price shocks by construction.

18.9.2 A Nonlinear Structural Model Allowing for Asymmetric Responses

As an alternative, Kilian and Vigfusson (2011a, 2011b) propose the structural model

$$\begin{aligned}
\Delta r_t &= \nu_1 + \sum_{i=1}^{p} b_{11,i}\Delta r_{t-i} + \sum_{i=1}^{p} b_{12,i}\Delta gdp_{t-i} + w_{1t},\\
\Delta gdp_t &= \nu_2 + \sum_{i=0}^{p} b_{21,i}\Delta r_{t-i} + \sum_{i=1}^{p} b_{22,i}\Delta gdp_{t-i}\\
&\quad + \textstyle\sum_{i=0}^{p} \gamma_i \Delta r_{t-i}^{net,+,3yr} + w_{2t},\\
\Delta r_t^{net,+,3yr} &\equiv \max(0, r_t - r_t^*),
\end{aligned} \tag{18.9.2}$$

where the structural shocks w_{1t} and w_{2t} are mutually uncorrelated and serially uncorrelated. The third equation defines the nonlinear regressors in the second equation. This model allows for all features that the censored oil price VAR model was intended to capture, while preserving the linearity of the equation of the real price of oil. It also directly imposes the conventional identifying assumption that the real price of oil is predetermined with respect to U.S. real GDP. Model (18.9.2) is not a conventional structural vector autoregression because the set of regressors differs across equations and because the net oil price increase affects Δgdp_t even in the impact period. As a result, it would not be possible to rewrite this model as a reduced-form model to be identified by a Cholesky decomposition. The model structure is nonlinear in that the effect of a positive oil price shock, w_{1t}, differs from the effect of a negative oil price shock of the same magnitude. Moreover, the response of real GDP changes disproportionately with the magnitude of the oil price shock and depends on the information set, Ω_{t-1}, at the time of the shock.

Model (18.9.2) encompasses as a special case the linear VAR(p) model

$$\begin{aligned}
\Delta r_t &= \nu_1 + \sum_{i=1}^{p} b_{11,i}\Delta r_{t-i} + \sum_{i=1}^{p} b_{12,i}\Delta gdp_{t-i} + w_{1t}\\
\Delta gdp_t &= \nu_2 + \sum_{i=0}^{p} b_{21,i}\Delta r_{t-i} + \sum_{i=1}^{p} b_{21,i}\Delta gdp_{t-i} + w_{2t},
\end{aligned} \tag{18.9.3}$$

for $\gamma_i = 0, i = 0, 1, \ldots, p$, in model (18.9.2).

In estimating model (18.9.2) we can take advantage of the fact that the model is linear in the parameters. The model may be estimated consistently using equation-by-equation least squares, and asymptotic inference on the parameters is standard. The inclusion of contemporaneous regressors in the second equation means that the structural shocks may be estimated directly. The construction of the structural impulse responses, in contrast, is complicated by the model being nonlinear in the variables which renders the impulse responses dependent on the state of the system at the time of the shock and the sign and size of the shock. It requires Monte Carlo integration, building on the algorithm outlined in Section 18.2.2.

18.9.3 Quantifying Nonlinear Responses to Oil Price Shocks

Having estimated the encompassing model (18.9.2) on the full sample, under the assumption of independent structural shocks, the structural impulse responses may be computed as follows:

Algorithm *Conditional response function at date t*

1. Consider a block of p consecutive values of Δr_τ, $\Delta r_\tau^{net,+,3yr}$, and Δgdp_τ for $\tau = t-p, \dots, t-1$. This sequence defines a history Ω_{t-1}. We are interested in quantifying the response of real GDP growth over the next $H+1$ quarters to an oil price innovation of magnitude δ occurring at date t, conditional on Ω_{t-1}.
2. We simulate two alternative time paths for Δr_{t+h} and Δgdp_{t+h} for $h = 0, \dots, H$ from model (18.9.2) by iterating the model forward, given Ω_{t-1} and given the estimated coefficients of model (18.9.2).
   In simulating future realizations of the data, we need to take a stand on the structural shocks, w_{t+h}, for $h = 0, \dots, H$. When generating the first time path, the value of w_{1t} is set equal to a prespecified value δ, denoting the magnitude of the oil price shock of interest. The realizations of the subsequent structural innovations, $w_{1,t+h}$ for $h = 1, \dots, H$, are drawn from the marginal empirical distribution of w_{1t}. The realizations of $w_{2,t+h}$ for $h = 0, 1, \dots, H$ are drawn independently from the marginal empirical distribution of w_{2t}.
   When generating the second time path, all $w_{1,t+h}$ and $w_{2,t+h}$ realizations for $h = 0, \dots, H$ are drawn independently from their respective marginal distributions.
3. We then calculate the difference between these two time paths for Δgdp_{t+h}, $h = 0, 1, \dots, H$.
4. We average this difference across m repetitions of steps 2 and 3, where m has to be large enough to invoke the law of large numbers.

This average represents the conditional response of Δgdp_{t+h} at horizon $h = 0, 1, \dots, H$ to a shock of magnitude δ given Ω_{t-1}, denoted by

$$I_{\Delta gdp}(H, \delta, \Omega_{t-1}).$$

The corresponding unconditional response function $I_{\Delta gdp}(H, \delta)$ for horizons $h = 1, \ldots, H$ is defined as the value of the conditional response function $I_{\Delta gdp}(H, \delta, \Omega^r)$ averaged over many randomly selected histories Ω^r, each of which is generated by randomly drawing with replacement a block of p consecutive observations from the observed data:

$$I_{\Delta gdp}(H, \delta) = \int I_{\Delta gdp}(H, \delta, \Omega^r) d\Omega^r.$$

The unconditional response, as implemented in Kilian and Vigfusson (2011a), is a measure of the overall effect of oil price shocks on U.S. real GDP. When studying the effects of oil price shocks during specific historical episodes commencing at date t, the appropriate impulse response measure instead is the conditional response to oil price shocks of magnitude δ, given the observed sequence of data preceding date t. The latter approach has been utilized in Kilian and Vigfusson (2017).

Bootstrap confidence intervals for the structural impulse responses may be constructed by bootstrapping the encompassing model. The asymptotic validity of this bootstrap approach has not yet been formally established, however.

18.9.4 Testing the Null of Unconditionally Symmetric Response Functions

One of the reasons researchers have investigated models based on censored oil price variables is the perception that the responses triggered by a positive and by a negative oil price shock of the same magnitude are asymmetric. In other words, the resulting impulse response functions are not perfect mirror images of one another. Typically, the expectation is that positive oil price shocks are more recessionary than negative oil price shocks are stimulating for the economy.

Symmetry in the impulse response functions requires that $\gamma_i = 0$ for $i = 0, 1, \ldots, p$. It may be tempting therefore to test for symmetry in the response functions by testing whether the γ_i parameters are jointly equal to zero. Such a test would not be informative about the degree of asymmetry in the impulse response functions, however, because the impulse responses are highly nonlinear functions of the slope parameters of model (18.9.2). It can be shown that, even when statistical tests reject the null of symmetric slope parameters, the impulse response functions may be nearly symmetric. As usual, a failure to reject the null of symmetric slopes may reflect the low power of the test and does not guarantee symmetric response functions.

Moreover, we know that the degree of asymmetry of the impulse response functions depends on δ. Slope parameter based tests are invariant to the magnitude of δ by construction and hence ill-suited for assessing the degree of asymmetry in the response functions. A more direct test of the symmetry of the response functions for a given δ involves three steps.

1. Estimate the encompassing model (18.9.2) on the full sample.
2. Compute $I_{\Delta gdp}(H, \delta)$ and $I_{\Delta gdp}(H, -\delta)$.
3. Test $\mathbb{H}_0 : I_{\Delta gdp}(H, \delta) = -I_{\Delta gdp}(H, -\delta)$ against the alternative $\mathbb{H}_1 : I_{\Delta gdp}(H, \delta) \neq -I_{\Delta gdp}(H, -\delta)$. In other words, under the null hypothesis the $(H + 1) \times 1$ vector $I_{\Delta gdp}(H, \delta) + I_{\Delta gdp}(H, -\delta)$ should be equal to a vector of zeros. Standardizing this vector based on a bootstrap estimator of the variance covariance matrix of the vector $I_{\Delta gdp}(H, \delta) + I_{\Delta gdp}(H, -\delta)$ results in the Wald test statistic

$$\left(\widehat{I}_{\Delta gdp}(H, \delta) + \widehat{I}_{\Delta gdp}(H, -\delta)\right)' \,\widehat{\Sigma}^{*-1}_{\widehat{I}^{*}_{\Delta gdp}(H,\delta)+\widehat{I}^{*}_{\Delta gdp}(H,-\delta)} \\ \times \left(\widehat{I}_{\Delta gdp}(H, \delta) + \widehat{I}_{\Delta gdp}(H, -\delta)\right),$$

where * denotes bootstrap estimators based on the unrestricted model (18.9.2). The null hypothesis of symmetric response functions is rejected if the Wald statistic exceeds the critical value from the $\chi^2(H + 1)$ distribution.

This procedure can be easily adapted to testing the symmetry of cumulative impulse response functions.

In practice, it is common to consider results for oil price shocks of size $\delta = \sigma$ and $\delta = 2\sigma$, where σ is one standard deviation of the estimated w_{1t}. When the test rejects the null of a symmetric response function at conventional significance levels, it is recommended to check the plots of the response functions to ascertain whether the rejection is also economically significant. In choosing H, it has to be kept in mind that for large H the $\chi^2(H + 1)$ approximation will become poor. In practice, a maximum horizon of one year is a common choice. Using one fixed H also avoids the problems associated with data mining across H.

18.9.5 Testing the Null of Conditionally Symmetric Response Functions

One concern with relying on tests of the null of unconditionally symmetric response functions is that the degree of asymmetry may evolve over time. In this case, unconditional tests may have low power against departures from symmetry. Kilian and Vigfusson (2017) propose a complementary test of the null of conditionally symmetric response functions on a specific date t. The test is conducted in three steps:

1. Estimate the encompassing model (18.9.2) on the full sample.
2. Compute $\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1})$ and $\widehat{I}_{\Delta gdp}(H, -\delta, \Omega_{t-1})$.

3. Test $\mathbb{H}_0 : I_{\Delta gdp}(H, \delta, \Omega_{t-1}) = -I_{\Delta gdp}(H, -\delta, \Omega_{t-1})$ against the alternative $\mathbb{H}_1 : I_{\Delta gdp}(H, \delta, \Omega_{t-1}) \neq -I_{\Delta gdp}(H, -\delta, \Omega_{t-1})$ using the Wald test statistic

$$\begin{aligned}&\left(\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1}) + \widehat{I}_{\Delta gdp}(H, -\delta, \Omega_{t-1})\right)' \\ &\quad \times \widehat{\Sigma}^{*-1}_{\widehat{I}^*_{\Delta gdp}(H,\delta,\Omega_{t-1})+\widehat{I}^*_{\Delta gdp}(H,-\delta,\Omega_{t-1})} \\ &\quad \times \left(\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1}) + \widehat{I}_{\Delta gdp}(H, -\delta, \Omega_{t-1})\right),\end{aligned}$$

where * denotes bootstrap estimators based on the unrestricted model (18.9.2). The null hypothesis of conditionally symmetric response functions is rejected if the test statistic exceeds the critical value from the $\chi^2(H+1)$ distribution.

18.9.6 Testing the Null of No Time Dependence

A common question is whether the conditional response is time dependent or whether it is indistinguishable from a linear VAR model response. Kilian and Vigfusson (2017) propose a test of the null of conditionally linear response functions on a specific date t. The test compares the conditional responses of real GDP growth to a real oil price shock from the nonlinear model, $\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1})$, to the corresponding quantities from a linear VAR model, denoted by $\widehat{I}^{VAR}_{\Delta gdp}(H, \delta)$. Note that the latter quantities do not depend on the specific date t for which the test is performed. There are four steps.

1. Estimate the encompassing model (18.9.2) and the linear VAR model (18.9.3) on the full sample.
2. Compute the conditional nonlinear response function $\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1})$ from the estimate of the encompassing model (18.9.2).
3. Compute the linear VAR response function $\widehat{I}^{VAR}_{\Delta gdp}(H, \delta)$ from the estimate of the linear VAR model (18.9.3).
4. Test $\mathbb{H}_0 : I_{\Delta gdp}(H, \delta, \Omega_{t-1}) = I^{VAR}_{\Delta gdp}(H, \delta)$ against the alternative $\mathbb{H}_1 : I_{\Delta gdp}(H, \delta, \Omega_{t-1}) \neq I^{VAR}_{\Delta gdp}(H, \delta)$ using the Wald test statistic

$$\begin{aligned}&\left(\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1}) - \widehat{I}^{VAR}_{\Delta gdp}(H, \delta)\right)' \\ &\quad \times \widehat{\Sigma}^{*-1}_{\widehat{I}^*_{\Delta gdp}(H,\delta,\Omega_{t-1})-\widehat{I}^{VAR*}_{\Delta gdp}(H,\delta)} \\ &\quad \times \left(\widehat{I}_{\Delta gdp}(H, \delta, \Omega_{t-1}) - \widehat{I}^{VAR}_{\Delta gdp}(H, \delta)\right),\end{aligned}$$

where * denotes bootstrap estimators. The null hypothesis of conditionally linear response functions is rejected if the test statistic exceeds the critical value from the $\chi^2(H+1)$ distribution. The

$\widehat{I}^{*}_{\Delta gdp}(H, \delta, \Omega_{t-1})$ and $\widehat{I}^{VAR*}_{\Delta gdp}(H, \delta)$ estimators are constructed by fitting models (18.9.2) and (18.9.3), respectively, to each of the bootstrap draws of the model data generated from the fitted encompassing model. The construction of $\widehat{\Sigma}^{*}$ takes account of the variances of the linear and the nonlinear bootstrap impulse response estimators as well as their covariances.

18.9.7 Conditional Prediction Error Decompositions

Impulse responses do not convey the cumulative effects of an entire sequence of oil price innovations on the log-level of U.S. real GDP over the course of a particular episode, starting at date t. Kilian and Vigfusson (2017) propose a conditional structural prediction error decomposition designed to address this problem. The total prediction error of model (18.9.2) at horizons h is $gdp_{t+h} - \mathbb{E}(gdp_{t+h}|\Omega_{t-1})$. A negative prediction error, for example, means that the model over-predicts U.S. real GDP. In other words, real GDP turned out lower than predicted.

One can decompose these prediction errors into the component explained by unforeseen oil price shocks, $w_{1,t+h}$, and the component explained by unforeseen residual shocks, $w_{2,t+h}$, in the second equation of the structural model. The conditional structural real GDP prediction error decomposition $D_{i,gdp}$, $i \in \{1, 2\}$, describes how real GDP would have evolved in the absence of further oil price innovations conditional on Ω_{t-1}:

$$D_{1,gdp}\left(h, \Omega_{t-1}\right) = \mathbb{E}\left(gdp_{t+h}|\left\{w_{1,t+i}\right\}_{i=0}^{h}, \Omega_{t-1}\right) - \mathbb{E}(gdp_{t+h}|\Omega_{t-1})$$

and, in the absence of further innovations in the second equation, conditional on Ω_{t-1},

$$D_{2,gdp}\left(h, \Omega_{t-1}\right) = \mathbb{E}\left(gdp_{t+h}|\left\{w_{2,t+i}\right\}_{i=0}^{h}, \Omega_{t-1}\right) - \mathbb{E}(gdp_{t+h}|\Omega_{t-1})$$

where $h = 0, \ldots, H$. The expectations must be evaluated by Monte Carlo simulation.

A similar analysis could also be conducted with the linear VAR model (18.9.3). Whereas in the linear model by construction

$$D_{1,gdp}\left(h, \Omega_{t-1}\right) + D_{2,gdp}\left(h, \Omega_{t-1}\right) = gdp_{t+h} - \mathbb{E}(gdp_{t+h}|\Omega_{t-1}), \tag{18.9.4}$$

in the nonlinear model (18.9.2) there may be interactions between the two structural innovations that prevent the decomposition from adding up. Such interactions often are negligible in practice, however, and can be checked based on equation (18.9.4).

18.9.8 Extensions

Applications of tests for symmetry to disaggregate data inevitably invite data mining as these tests are applied repeatedly (see Inoue and Kilian 2004). Herrera, Lagalo, and Wada (2011) extend the Kilian-Vigfusson methodology by using a bootstrap approach to generate data-mining robust critical values under the null of symmetric response functions. Kilian and Vigfusson (2017) discuss extensions to nonlinear structural models with additional variables. An extension of the Kilian-Vigfusson methodology to structural vector error correction models is developed in Venditti (2013). Although the example of oil price shocks is well suited to illustrating the econometric issues in dealing with censored variables, the applicability of this methodology is not limited to studying the effects of oil price shocks. For example, Hussain and Malik (2016) use this methodology to investigate asymmetries in the response of U.S. real GDP to fiscal policy shocks.