CHAPTER 4

Shocks, Sign Restrictions, and Identification

Harald Uhlig

This chapter highlights some key issues in the use of sign restrictions for the purpose of identifying shocks. It does so by examining two benchmark examples. In the first part, I discuss a generic example of demand and supply, seeking to identify a supply shock from price–quantity data. In the second part, I discuss a generic example of Bayesian vector autoregressions and the identification of a monetary supply shock. Along the way, I formulate some principles and present my view on some of the recent discussion and literature regarding sign restrictions.

1 INTRODUCTION

The approach of sign restrictions in time series analysis has generated an active literature, many successful applications, and a lively debate. The procedures are increasingly easy to use, with implementations in econometric software packages such as RATS or with ready-to-implement code in a variety of programming languages; see, e.g., Danne (2015) as one example. Let me say from the outset that I am very happy about that, including those contributions that have criticized my own work, sometimes sharply. Skepticism and critique is crucial for science to advance, so all power to them! That should not prevent me from critiquing back, of course, and that is partly what this chapter will be about. Debate is good.

While Leamer (1981) surely deserves being highlighted here, I believe that the literature pretty much started with Dwyer (1998), Faust (1998) and its discussion, Uhlig (1998), Canova and Pina (1999), Canova and de Nicolo (2002),

This research has been supported by the NSF grant SES-1227280 and the INET grant #INO1100049, "Understanding Macroeconomic Fragility". I have an ongoing consulting relationship with a Federal Reserve Bank, the Bundesbank and the ECB. I am grateful to Alexander Kriwoluzky for the encouragement to use the material of the first main section concerning the identification of supply and demand as the topic of this chapter.

as well as my "agnostic identification" paper Uhlig (2005b). This one was published quite a number of years after my discussion of the Faust paper, but that discussion shows that I had already developed my methodology then, and that imposing sign restrictions on impulse responses and not just on impact can add considerable bite. There are deep connections to the seemingly different literature on partial identification and estimation subject to inequality restrictions: rather than review that literature, let me just point the reader to the excellent discussions on this topic by Canay and Shaikh (2017) as well as by Ho and Rosen (2017), appearing elsewhere in this volume, or, say, Kline and Tamer (2016).

The purpose of this chapter instead is to help shed some light on some issues that arise when using sign restrictions. From discussions with fellow researchers or occasional statements in the existing literature, I find that some of these issues can be a cause of confusion, misunderstanding, misinterpretation or misjudgement. I feel that these issues should be given some thought: so here are mine. I will list my main lessons as "principles." That surely sounds more grandiose than is intended: perhaps "recommendation" or "my current random thought on this issue" would be a better label, but let me proceed with the more pompous label without further apology. Other researchers might well disagree with some or many of them. Let the discussions commence.

I will organize my discussion around two key questions:

1. What happens after a supply shock?
2. What happens after a monetary policy shock?

I provide some generic econometric perspective as well as occasionally engaging with some of the recent debate and literature concerning sign restrictions. Of course, both of these questions have been part of the bread-and-butter econometrics literature for years. Since much (perhaps too much?) has been written about them, and since much of that might be familiar to the reader, they are thus particularly useful to discuss the issues at hand. If you do not feel familiar with that earlier literature at all, fear not, though. This chapter is largely self-contained.

The first question concerns one of the most generic questions one may wish to ask, given market observations on quantities and prices. This question is also, in essence, at the core of many other applications of the sign restriction methodology and, in fact, identification generally: how should one disentangle contemporaneous observations into their underlying stochastic sources? How can one identify the individual shocks moving the data? For the first question, I shall disregard the time series perspective entirely and focus solely on the identification issue. Much has already been written about the baseline case of interpreting price–quantity data in terms of their underlying movements in supply and demand, and using that baseline case for a deeper understanding of more general issues. Some important examples of that literature are Leamer (1981) as perhaps the first author to emphasize the importance of inequality constraints, Angrist, Graddy, and Imbens (2000) and, more recently,

Baumeister and Hamilton (2015) with a focus on sign restriction identification in time series analysis. The first part of this chapter is based on my draft paper, Uhlig (2005a), considerably revised and updated for the purpose here. Compared to the other two papers, I will therefore introduce yet another way of mathematically framing the issue, with all due apologies. There are deep connections between them all, of course, and I will be able to only superficially touch upon them.

For better or worse, the second question has become the showcase example for much of the time series literature, analyzing macroeconomic interrelationships with the aid of vector autoregressions. Here, the time series perspective is of considerable importance, and provides me with the opportunity to expose some key issues arising. The previous issues of identification have not gone away, however: far from it.

2 SUPPLY AND DEMAND

Consider Figure 1. Let's say it shows an artificially generated sample of draws $((P_i, Q_i))_{i=1}^n$ of equilibrium price–quantity pairs, from some repeated market-clearing observations of some market subject to some *i.i.d.* shocks. I have removed the mean: thus, the points cluster around $(0, 0)$ rather than some average price and quantity. One may wish to add those back in or also consider them to be part of the challenge (or information) of estimating supply and demand, but I shall proceed with abstracting from that. I seek to understand how the equilibrium shifts in the wake of a shock to supply.

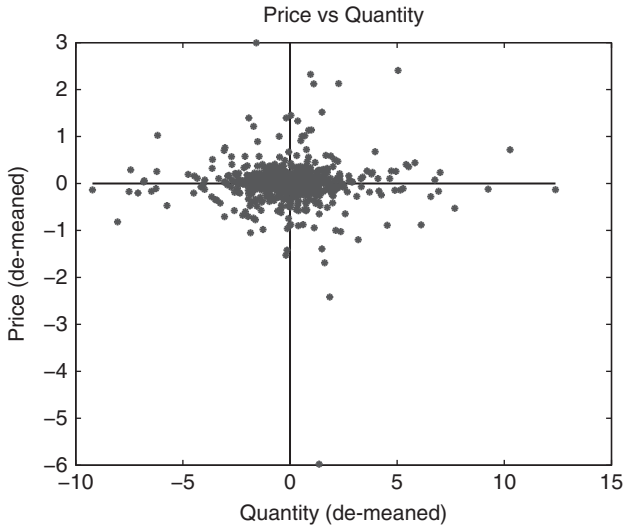

Figure 1 Some artificial sample of draws of equilibrium price–quantity pairs, de-meaned

Let me imagine, thus, that the price–quantity pairs shown in Figure 1 arise as the intersection of supply and demand curves with constant slopes, where the curves are shifted by random shocks, but would intersect at $P = 0, Q = 0$ without such shocks. If you wish, you may interpret the price P and the quantity Q here on a log scale, relative to some mean, so that the slopes of the supply and demand curves represent the elasticities of supply and demand. I shall restrict these shocks to be *i.i.d.* across observations for the purpose of the discussion here. It should be clear that considerable structure already has been imposed on the problem.

To interpret the data then, one would like to know the slopes of these supply and demand curves, and disentangle the equilibrium observations (P_i, Q_i) into their underlying supply and demand shocks. As is well known, this gives rise to a challenging identification problem. The raw price–quantity data, as shown in Figure 1, is insufficient to identify the slopes of demand and supply. Figure 2 exhibits the challenge. There, I picked one of the (P_i, Q_i) observations and show, in two different ways, how this observation could have arisen. In the panel on the left, the supply and demand curves are fairly steep, whereas they are quite flat in the panel on the right. Both possibilities are consistent with the chosen (P_i, Q_i) observation. I will soon give this statement a bit more mathematical structure.

It should be clear then that additional identifying restrictions are needed. But which? That then is the key issue: which restrictions should one impose?

Figure 3 shows one of many possibilities. There, the condition is imposed that demand is vertical, i.e., completely price-inelastic. As a consequence, supply shocks move prices, not quantities. We now know how far the demand curve must have been moved by the demand shock: exactly as much so as to move the demand curve to the observed quantity rather than zero. So far, we still do not know the slope of the demand curve: it should be clear, from inspection of the picture, that there still is a range of possibilities. But nonetheless, some progress has been made.

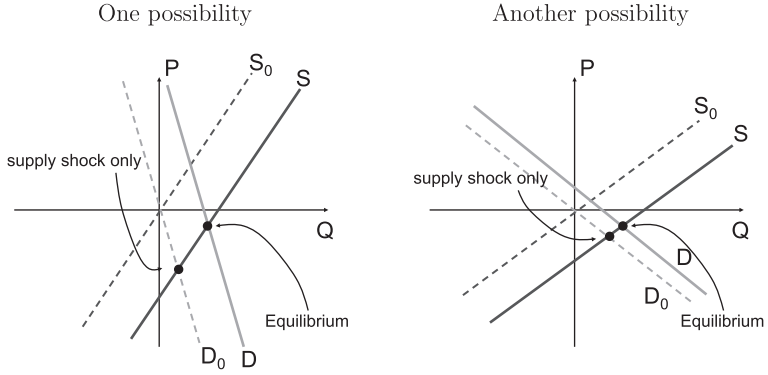

Figure 2 The identification problem: two examples for supply and demand shocks giving rise to the same observed price–quantity pair

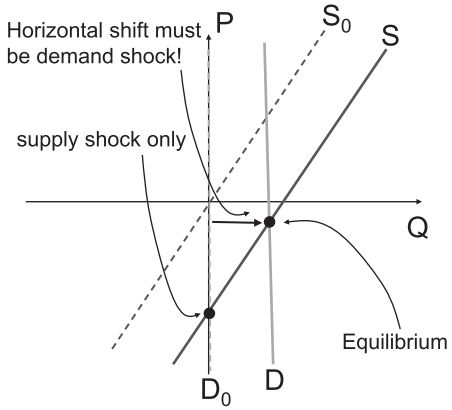

Figure 3 Identifying restriction: demand is vertical, i.e., completely price-inelastic. As a consequence, supply shocks move prices, not quantities.

Is the restriction of vertical demand a good one? It may be, under certain circumstances. Indeed, in many macroeconomic time series applications of identified vector autoregressions, it is quite popular to bring the series at hand into some order and to then to employ a Cholesky decomposition to identify their shocks. That decomposition assumes that shocks to one series only influence that series and series that come later in that order, but not those that come earlier. In the example here, one would order the data with quantity coming first and prices coming second. A shock to quantity would then be allowed to move both price and quantities and would correspond to a demand shock. A shock to prices would then be interpreted as a supply shock and assumed to not also move quantities, per the Cholesky decomposition, thus implementing the vertical demand curve assumption.

So, is this a good idea or not? My answer may be rather obvious. If the researcher can defend this causal ordering, sure. If not, then not. Let me formulate this as my first two principles. They may sound rather banal, but they are central in almost all applied work, and worth reflecting on in any particular application.

Principle 1: *If you know it, impose it!*

Principle 2: *If you do not know it, do not impose it!*

Of course, it is good to impose what you know: it allows you to deduce more from the data. There is usually no good reason to discard such information. Some of the criticism of the sign restriction seems to imply that the sign restriction literature tended to encourage users to disregard more precise identifying assumptions: that, of course, is not the case.

The astute reader may be asking, though, what I mean by "knowing." Do we really know anything at all? That would make it impossible ever to apply

Principle 1. I suggest interpreting the treshold much more liberally and with a practical mind. McCloskey (1998) has argued that economics is rhetoric. One does not need to go as far as McCloskey to realize that the main objective of research findings is nonetheless indeed often to move and, ideally, convince skeptical audiences and colleagues. So, if some assumption can reasonably be imposed to that end, go ahead. If not, then be careful.

2.1 Sign Restrictions

Are there assumptions in the situation at hand, then, that are palatable to most audiences? I shall argue that there are. Supply slopes up. Demand slopes down. Always? Perhaps not. But usually, typically. As assumptions go, these may be among the most agreeable to impose. Let me give this a bit more structure: another reasonable assumption will then emerge.

The slopes of the supply and demand curves are key. Since slopes can be horizontal or vertical, it seems most natural to me to characterize the slope by the point that these curves intersect with the unit circle in (P, Q)-space. Likewise, it seems most natural to me to then characterize shifts by moving those curves in a direction orthogonal to their slopes. Additionally, this turns out to be algebraically convenient. Since we are shifting the entire curve rather than worrying how to shift a particular point on the supply curve, it really is immaterial whether one assumes the random shift to be exactly in an orthogonal direction or in a somewhat aligned fashion. The orthogonality assumption should be understood to be a convenient normalization. Obviously, other parameterizations are possible.

With that, the supply curve is the line

$$\begin{bmatrix} P \\ Q \end{bmatrix} = z_S \lambda_S + x_S \sigma_S \epsilon_S, \ \lambda_S \in \mathbf{R} \tag{1}$$

where z_S is the direction of the supply curve, x_S is the direction of supply curve shift, and $\sigma_S \epsilon_S$ is the supply shock, where ϵ_S is normalized to have unit variance. I normalize z_S to be of unit length and x_S to be orthogonal to z_S,

$$z_S = \begin{bmatrix} \cos(\nu_S) \\ \sin(\nu_S) \end{bmatrix}, \ x_S = \begin{bmatrix} -\sin(\nu_S) \\ \cos(\nu_S) \end{bmatrix} \tag{2}$$

where $\nu_S \in [0, 2\pi)$ and where $\sigma_S \geq 0$.

In most cases, it is then reasonable to impose the following assumption.

Assumption A.1 *Supply slopes upward*

I impose the additional normalization that a positive supply shift is towards higher Q. Such normalizations are often innocous, but sometimes they are not, especially in a time series context. For example, if small quantity movements now are followed by larger movements later on, which additionally show a more stable sign, it may be much more sensible to normalize shifts by the subsequent later movement rather than by the contemporaneous movement. I

once witnessed a presentation by leading researchers who tripped themselves up badly by normalizing by the sign of the initial response instead. Here, of course, there is no time series aspect to worry about. Together with Assumption A.1 and $\nu_S \in [0, 2\pi)$, we then have more formally that

Assumption A.2 *Supply slopes upward and a positive supply shift is towards higher Q:* $\nu_S \in [0, \pi/2]$.

I shall additionally impose the distributional assumption

Assumption A.3

$$\epsilon_S \sim \mathcal{N}(0, 1). \tag{3}$$

While a standard deviation of 1 is simply a normalization, the normal-distribution assumption may be too much, in some circumstances and for some audiences. I read it as a convenient benchmark for making further progress with a more formal statistical analysis.

It is also here where the normalization of a shift as a shift in an orthogonal direction compared to the original curve matters. Another possibility would be to parameterize the shift by the quantity change, for a fixed price change. This is pursued by Baumeister and Hamilton (2015), who show that unappealing distributions such as Cauchy distributions may result. It is clear where their result comes from. For very flat supply curves, large quantity changes are then needed in order to generate a given price response: as the supply curve becomes completely flat asymptotically, the required quantity change diverges to infinity. There may be circumstances to proceed in this manner. Sometimes the focus question is indeed about the effect of supply shocks with a given price change. An example might be to ask about the impact of a monetary policy shock, when interest rates are raised by 100 basis points; see also Figure 10. At that point, though, it may be useful to contemplate whether this normalization is truly at the heart of the question or just a convenience for phrasing the question. It may well be the latter, in which case another normalization would be perfectly legitimate. Conversely, if one insists on fixing the impact on one variable such as the price, it might then be appropriate to bound the range of supply curves per bounding by the range of quantity reactions. Though not used much, it should not be hard to convince an audience that imposing an upper bound on, say, the quantity reaction is legitimate. Such a bound then avoids the unappealing properties of the distributions documented by Baumeister and Hamilton (2015).

Likewise, the demand curve is the line

$$\begin{bmatrix} P \\ Q \end{bmatrix} = z_D \lambda_D + x_D \sigma_D \epsilon_D, \ \lambda_D \in \mathbf{R} \tag{4}$$

where $\sigma_D \geq 0$, where

$$z_D = \begin{bmatrix} \cos(\nu_D) \\ \sin(\nu_D) \end{bmatrix}, \; x_D = \begin{bmatrix} \sin(\nu_D) \\ -\cos(\nu_D) \end{bmatrix} \tag{5}$$

and where $\nu_D \in [0, 2\pi)$. In parallel with Assumption A.1, it is usually reasonable to impose

Assumption A.4 *Demand slopes downward.*

Again, we add the normalization that a positive demand shift is towards higher Q. In sum, we shall impose

Assumption A.5 *Demand slopes downward and a positive demand shift is towards higher Q:* $\nu_D \in [\pi/2, \pi]$.

Furthermore, I shall once again impose the distributional assumption that

Assumption A.6

$$\epsilon_D \sim \mathcal{N}(0, 1). \tag{6}$$

While the distributional assumptions may impose too much structure, the following assumption should often be broadly agreeable even without it, however.

Assumption A.7 ϵ_S *and* ϵ_D *are mutually independent.*

This is the third assumption promised above. Without that assumption, it becomes conceptually murky what one means, when analyzing equilibrium properties in response to a supply shock. If the supply shock hits, does the demand shock move too? Or is it the other way around: should one first cleanse the supply shock of the movement by the demand shock? A chicken-and-egg problem arises. Assumption A.7 may arguably impose too much in certain circumstances, however. Perhaps, a third source of randomness is moving both. But then, one ought to find it and model it! It would then not be legitimate to call the supply shock a "shock": rather, it should be decomposed into a response to stochastic movements elsewhere and a genuine own-shock component. With that, let me announce my next principle.

Principle 3: *Shocks are independent.*

Notationally, it is convenient to summarize the sign restrictions in Assumptions A.2 and A.5 by "+" or "−" or "?" (in case of no restriction) to indicate the sign of the response to a positive shock. In order to learn about equilibrium behavior in response to supply shocks, it may be tempting to just impose the sign restrictions of Assumption A.2 on supply shocks. Furthermore, in larger systems, one may not feel comfortable imposing sign restrictions on other shocks as well. In the context here, though, it seems natural to impose sign

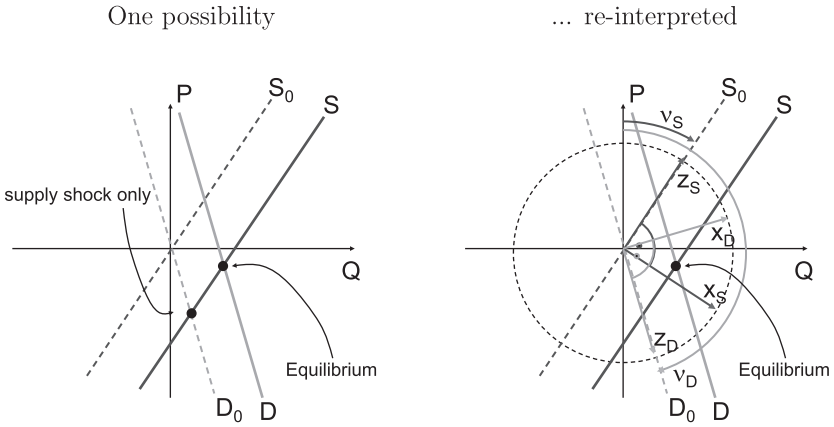

Figure 4 A graphical representation of the mathematical definitions for the slopes and shifts of demand and supply

restrictions on both supply and demand shocks, i.e., to impose both Assumptions A.2 and A.5. I shall investigate the consequences of both choices. If only sign restrictions on supply shocks are imposed, we have

$$\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{array}{c} \begin{matrix} \epsilon_S & \epsilon_D \end{matrix} \\ \begin{bmatrix} - & ? \\ + & ? \end{bmatrix} \end{array}. \tag{7}$$

If sign restrictions on both, supply and demand shocks, are imposed, we have

$$\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{array}{c} \begin{matrix} \epsilon_S & \epsilon_D \end{matrix} \\ \begin{bmatrix} - & + \\ + & + \end{bmatrix} \end{array}. \tag{8}$$

I can now bring all that structure to bear on the decompositions shown above in the left panel of Figure 2. Figure 4 provides a graphical representation of that additional structure.

Statistically, I can now proceed as follows. The structural parameters are

$$\theta = (\nu_S, \nu_D, \sigma_S, \sigma_D). \tag{9}$$

Note that θ is four-dimensional. Further, and as a rule, it is typically reasonable to impose the assumption that

Assumption A.8 $\sigma_S > 0$ *and* $\sigma_D > 0$.

Without it, we would reduce the supply-and-demand structure to one with a single shock or no shock at all. One should note that there are important exceptions to this rule, though. Simple real business cycle models are sometimes driven by a few shocks only. If more time series are used to estimate such models than there are available shocks, this creates singularity problems.

It is then best to introduce sufficiently many shocks, either as measurement error or, often better, as a structural part of the model, to avoid these singularities. One needs to realize, though, that the introduction of these additional shocks will then modify or constrain how the data are interpreted in terms of the original shocks: different shock additions can create potentially quite different interpretations. With that caveat, I shall announce

Principle 4: *Have at least as many shocks as the length of the data vector.*

In the context of our price–quantity example, that means we ought to have two shocks, i.e., Assumption A.8.

The following are some convenient definitions

$$
\begin{aligned}
Z &= \begin{bmatrix} z_D & z_S \end{bmatrix} = \begin{bmatrix} \cos(\nu_D) & \cos(\nu_S) \\ \sin(\nu_D) & \sin(\nu_S) \end{bmatrix} \\
X &= \begin{bmatrix} x_S' \\ x_D' \end{bmatrix} = \begin{bmatrix} -\sin(\nu_S) & \cos(\nu_S) \\ \sin(\nu_D) & -\cos(\nu_D) \end{bmatrix} \\
\Omega &= \begin{bmatrix} \sigma_S & 0 \\ 0 & \sigma_D \end{bmatrix} \\
\epsilon &= \begin{bmatrix} \epsilon_S \\ \epsilon_D \end{bmatrix}
\end{aligned}
$$

Given ϵ_S and ϵ_D, one can solve for λ_S and λ_D as well as P and Q, such that the supply function and the demand function both deliver the price–quantity pair (P, Q), i.e., so that supply and demand intersect and equilibrium is achieved:

Proposition 1 *Equilibrium, i.e., equality of demand and supply is given by*

$$
\begin{bmatrix} P \\ Q \end{bmatrix} = X^{-1}\Omega\epsilon. \tag{10}
$$

Proof "Guess and verify". I need to show that there is a solution λ_D and λ_S, generating this price–quantity pair. For supply, substitute (10) for the left-hand side in (1). Multiply with X to find

$$
\begin{bmatrix} \sigma_S\epsilon_S \\ \sigma_D\epsilon_D \end{bmatrix} = \begin{bmatrix} 0 \\ x_D' z_S \end{bmatrix} \lambda_S + \begin{bmatrix} 1 \\ x_D' x_S \end{bmatrix} \sigma_S\epsilon_S. \tag{11}
$$

Proceed likewise for demand. The first row of (11) and the second row of the similar equation for demand together imply that

$$
X \begin{bmatrix} P \\ Q \end{bmatrix} = \Omega\epsilon \tag{12}
$$

which implies (10). If one wishes to proceed further, note that the second row of (11) can be solved for

$$
\lambda_S = (\sigma_D\epsilon_D - x_D' x_S \sigma_S\epsilon_S)/(x_D' z_S)
$$

noting that $x_D' z_S \neq 0$ per $(\nu_S - \nu_D) mod\, \pi \neq 0$, i.e., per ruling out parallel demand and supply curve. The solution for

$$\lambda_D = (\sigma_S \epsilon_S - x_S' x_D \sigma_D \epsilon_D)/(x_S' z_D)$$

follows likewise. □

This allows me to characterize the statistical distribution of the price–quantity data, exploiting the normal distribution assumption for the supply and demand shock. Given θ,

$$\begin{bmatrix} P \\ Q \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma(\theta)\right)$$

where

$$\Sigma(\theta) = A(\theta)A(\theta)' \tag{13}$$

with

$$A(\theta) = X^{-1}\Omega = \frac{1}{x_D' z_S}\begin{bmatrix} \sigma_S z_D & \sigma_D z_S \end{bmatrix} =: [a_S, a_D] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{14}$$

The matrix $A(\theta)$ has a natural interpretation, which is important to the lead question at hand. Given θ, the first column of $A(\theta)$ is the movements in price and quantity $[P, Q]'$ after a positive supply shock one standard deviation in size. Such a shock will lead to a movement along the demand curve, i.e., in the direction z_D. The length of the movement is given by $\sigma_S/(x_D' z_S)$ for a unit-sized supply shock, and therefore depends on the standard deviation σ_S of the non-normalized supply shock as well as the angle between the supply and demand curves, characterized by the inner product $x_D' z_S$. Likewise, the second column of $A(\theta)$ is the movements in price and quantity $[P, Q]'$ after a positive demand shock one standard deviation in size. As per our conventions stated above, we shall impose the normalization that the bottom row is positive, so that positive demand or supply shocks have a positive impact on quantities. For any given decomposition in (13) one can always achieve this by possibly flipping the signs of the columns of $A(\theta)$. As a consequence, $sgn\,(\det(A(\theta))) = -1$. Note furthermore that the mapping $\theta \mapsto A(\theta)$ is invertible, due to Assumption A.8.

Given finite data on pairs of prices and quantities, one can estimate $\Sigma(\theta)$. There are finite-sample issues in doing so, but they are not of particular relevance for the discussion at hand. Let us then assume that we see "enough" data so that we learn

$$\Sigma(\theta) = \Sigma = E\left[\begin{bmatrix} P \\ Q \end{bmatrix}\begin{bmatrix} P & Q \end{bmatrix}\right]$$

(assuming that the means of P and Q are zero) perfectly. It should also be clear that there is not more to learn from the data. $\Sigma(\theta)$ is it.

I can now describe the identification problem more succinctly. The parameter vector θ is four-dimensional. The variance-covariance matrix $\Sigma(\theta)$ is three-dimensional, due to symmetry. There is therefore one degree of freedom, in decomposing $\Sigma(\theta)$ into $A(\theta)$ and $A(\theta)'$ in (13), aside from some convention regarding the sign of the columns.

One additional exact restriction can render this problem exactly identified. For example, imposing that demand is vertical per $\nu_D = \pi$, as shown in Figure 3 will do the trick. This may actually be odd: wasn't it the case that one still could not know the slope of the supply function in Figure 3? How come we know it now? The reason is Principle 3 or, more precisely, the assumed independence of our normally distributed shocks ϵ_S and ϵ_D. Now, the observed covariance of prices and quantities can be ascribed to vertical demand and mutually independent shocks ϵ_S and ϵ_D only for one particular slope of the supply curve, as the calculations above show.

So, one additional exact restriction would be wonderful. But generally, it is hard to come by. Is it really reasonable to impose the condition that demand is vertical, say? As I discussed above: perhaps not. And perhaps then, we only have the sign restrictions, as embodied by Assumptions A.2 and A.5. Let me examine the consequences. The sign restrictions deliver inequalities on the slope parameters for the demand and supply curves, and therefore deliver sets Θ of θs compatible with the data. Some are excluded, but plenty are still left that merit consideration.

If we restrict only demand to be downward sloping, i.e., if we restrict supply shocks to move P, Q in opposite direction, then

$$\Theta_D = \{\theta = (\nu_S, \nu_D, \sigma_S, \sigma_D) \mid \nu_D \in [\pi/2, \pi], \nu_S \in (\nu_D - \pi, \nu_D), \sigma_S > 0, \sigma_D > 0\}.$$

If we restrict demand and supply, then

$$\Theta_{DS} = \{\theta = (\nu_S, \nu_D, \sigma_S, \sigma_D) \mid \nu_S \in [0, \pi/2], \nu_D \in [\pi/2, \pi], (\nu_S, \nu_D) \neq (0, \pi), \sigma_S > 0, \sigma_D > 0\}.$$

So, if sign restrictions are all we have, then in the example at hand, we get the next observation:

Principle 5: *Without an additional exact restriction, the sign restrictions deliver a one-dimensional set Θ of θs, which all could have generated the data.*

Note that this is true even asymptotically, i.e., if Σ is known exactly. Note also that one could parameterize this resulting set Θ by, say, $\nu_s \in N = [\underline{\nu}_S, \bar{\nu}_S]$.

How useful are sign restrictions? To think about this further, let me introduce another way to write the parameter vector θ and to parameterize the set Θ resulting from sign restriction. Consider the Cholesky decomposition

$$\Sigma = LL'$$

of Σ into the product of a lower triangular matrix L with its own transpose, normalized to have positive diagonal entries. Consider the decomposition of Σ shown in equation (13). Any such decomposition can be written as

$$LR = A(\theta) \tag{15}$$

for some orthogonal matrix

$$R = \begin{bmatrix} \cos(\mu) & \sin(\mu) \\ \sin(\mu) & -\cos(\mu) \end{bmatrix} \tag{16}$$

and some $\mu = \mu(\theta)$. Conversely, any μ and consequently R as in (16) generates a candidate decomposition in equation (13), provided $A_{21} \geq 0$ and $A_{22} \geq 0$. A satisfies the sign restrictions on demand and supply, iff the signs as indicated in equations (8) are satisfied, i.e., if

$$A_{11} \leq 0, A_{21} \geq 0, A_{12} \geq 0, A_{22} \geq 0$$

Likewise, A satisfies the sign restrictions on supply shocks, iff

$$A_{11} \leq 0, A_{21} \geq 0$$

as indicated in equation (7) Equation (10) shows, how to back out the underlying shocks, given knowledge of $A(\theta)$:

$$\begin{bmatrix} \epsilon_S \\ \epsilon_D \end{bmatrix} = (A(\theta))^{-1} \begin{bmatrix} P \\ Q \end{bmatrix} = RL^{-1} \begin{bmatrix} P \\ Q \end{bmatrix}, \tag{17}$$

noting that L^{-1} is known with Σ and that $R = R^{-1}$.

With that, we can re-parameterize θ. Write

$$\Sigma = \begin{bmatrix} \sigma_P^2 & \sin(\phi)\sigma_P\sigma_Q \\ \sin(\phi)\sigma_P\sigma_Q & \sigma_Q^2 \end{bmatrix}, \quad L = \begin{bmatrix} \sigma_P & 0 \\ \sin(\phi)\sigma_P & cos(\phi)\sigma_Q \end{bmatrix}$$

for some $\phi \in (-\pi, \pi)$. Note that $\sin(\phi)$ is the correlation between price and quantity. For example, $\phi < 0$ iff $\text{corr}P, Q < 0$. Thus, instead of $\theta = (\nu_S, \nu_D, \sigma_S, \sigma_D)$, let me use $\psi = (\phi, \mu, \sigma_P, \sigma_Q)$. Calculate, that

$$A = LR = \begin{bmatrix} a_S(\psi) & a_D(\psi) \end{bmatrix}$$

where now

$$a_S(\psi) = \begin{bmatrix} \cos(\mu)\sigma_P \\ \sin(\phi + \mu)\sigma_Q \end{bmatrix}, \quad a_D(\psi) = \begin{bmatrix} \sin(\mu)\sigma_P \\ -\cos(\phi + \mu)\sigma_Q \end{bmatrix}. \tag{18}$$

This provides us with another way to look at the mathematical structure shown in the right panel of Figure 4, exploiting the sign structure as indicated in (8) to constrain μ, given ϕ. The left panel of Figure 5 shows the consequences of imposing the supply sign restriction of Assumption A.2 or of the first column of (8) in the two-dimensional space of R_{11} and R_{12}. This is the first row of R: with that, we know the second row of R per (16). Likewise, the right

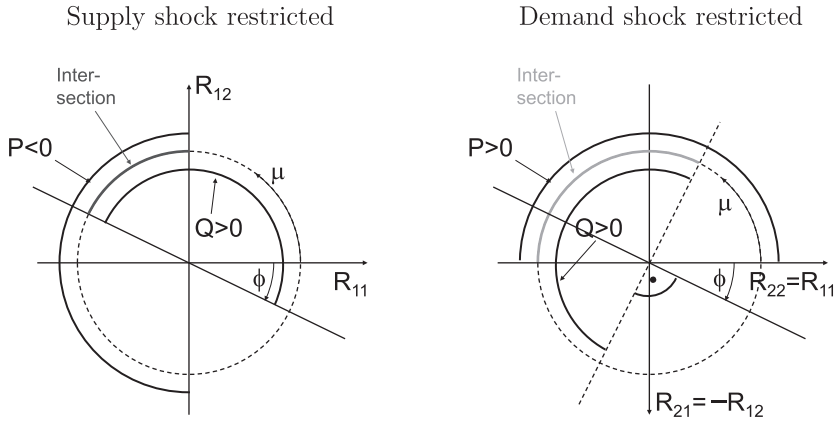

Figure 5 Sign restrictions and their consequences for μ, ϕ and R

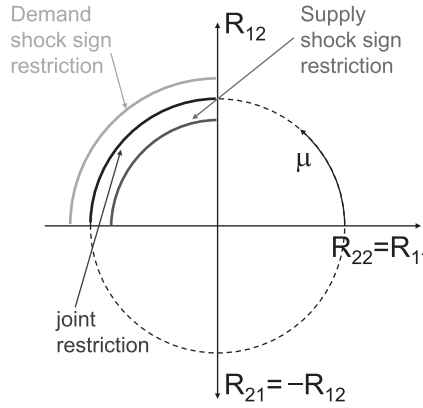

Figure 6 Sign restrictions, if $\text{corr}(P, Q) = 0$

panel of Figure 5 shows the consequences of imposing the demand sign restriction of Assumption A.5 or of the second column of (8) in the two-dimensional space of R_{11} and R_{12}.

Figures 6, 7, and 8 now show the consequences of imposing both sign restrictions, for different cases of observed price–quantity correlations. The worst case scenario surely is 6 of zero correlation: in that case, the joint restriction is the same that results from just imposing sign restrictions on demand or just sign restrictions on supply. There simply is not much one can learn about the slopes of demand and supply in this case. With an uncorrelated cloud of price and quantity, one might be tempted to draw diagonal supply and demand curves through them, but they might as well be very close to horizontal or vertical. One simply cannot tell. Matters change, once correlation between prices and quantities is observed. Figure 7 shows what happens for

$\text{corr}(P, Q) > 0$ $\text{corr}(P, Q) \to 1$

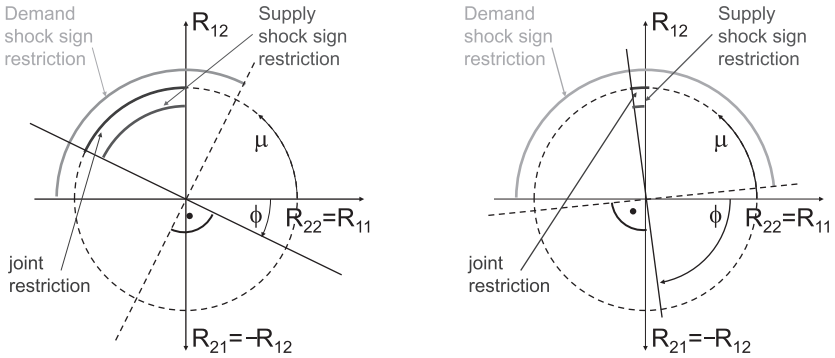

Figure 7 Sign restrictions on demand and supply, and their consequences for μ, ϕ and R, if prices and quantities are positively correlated. As $\text{corr}(P, Q) \to 1$, the supply shock sign restriction suffices for asymptotically exact identification.

$\text{corr}(P, Q) < 0$ $\text{corr}(P, Q) \to -1$

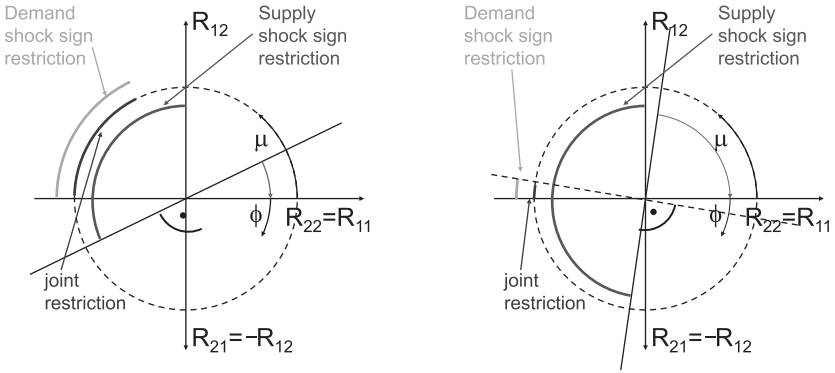

Figure 8 Sign restrictions on demand and supply, and their consequences for μ, ϕ and R, if prices and quantities are negatively correlated. As $\text{corr}(P, Q) \to -1$, the demand shock sign restriction suffices for asymptotically exact identification.

positive correlation between prices and quantities. The supply sign restriction is now the one that imposes more bite, whereas the demand sign restriction becomes less restrictive. In the limit, as $\text{corr}(P, Q) \to 1$, the supply shock sign restriction suffices for asymptotically exact identification. Given that we have imposed the independence of supply and demand shocks, a nearly perfectly positive alignment of price–quantity pairs must mean that we have found the supply curve, provided we are willing to impose Assumption A.1, that

the supply curve indeed slopes up. Once we know the supply curve, we have found one additional parameter restriction and identification is achieved. Likewise, Figure 8 shows what happens for negative correlation between prices and quantities, where the demand sign restriction is now the one that imposes more bite and achieves exact identification in the asymptotic limit, as the correlation becomes perfectly negative, while the supply sign restriction becomes increasingly useless.

Let me formulate the lesson from these observations:

Principle 6: *Sign-restricting both shocks compared to just sign-restricting one shock can help: sometimes a lot, sometimes not at all.*

2.2 Inference

What are practical ways to generate the set Θ of θs or, equivalently, the sets M of μs, which are consistent with the imposed sign restrictions? One may go ahead and proceed algebraically, as described above. This should work well in two dimensions, but will be increasingly challenging in more than two dimensions. The following draw-and-reject procedure is then more straightforward and of use in many situations:

1. Draw μ from prior on $[0, 2\pi]$.
2. Calculate the resulting $A = LR(\mu)$.
3. Check sign restrictions:
   (a) If satisfied, keep draw.
   (b) If not satisfied, reject draw.

If one rejects many draws, one may feel that something is going wrong. But here, the opposite is really true. Consider the right panels in Figures 7 or 8, when imposing both sign restrictions. The closer the correlation is to the extremes and thus, the sharper the identification, the smaller the range of μ that results in non-rejection. This is an important insight that is often misunderstood.

Principle 7: *When a lot of draws are rejected, the identification is sharp. Good!*

The judgement "good" is not meant to say "good" for the patience of the researcher, waiting for results, or "good" for whoever has to pay the electricity bill from running the computer at full capacity for a long time. "Good" here means that it is good in terms of being able to learn something from the data: that, after all, is what we normally care about the most.

So far, we have ignored the issue of finite samples, but it is time to return to it and its practical implications. Here, the question or object of interest is quite important. One can treat the vector θ as the parameter of interest or the

entire set Θ. That is, while some researchers may be interested in estimating $\hat{\theta}$ or $\hat{\mu}$ or provide probabilistic statements about it, other researchers may be interested in $\hat{\Theta}$ or $\hat{M}$ and probabilistic statements about that. There is no right or wrong here. The question of interest is chosen by the researcher, and the econometric procedures are there to help answer it. My own research has so far entirely been focused on treating θ or μ as the object of interest, but there is an interesting and developing literature on set identification focusing, effectively, on the set Θ or M instead; see, e.g., the discussions on this topic by Canay and Shaikh (2017) as well as by Ho and Rosen (2017), appearing elsewhere in this volume, or, say, Kline and Tamer (2016). These procedures then may give different answers, but that should not surprise us at all. Let me formulate the following rather obvious principle, which is sometimes oddly ignored in debates on this topic (or other topics, for that matter).

Principle 8: *Different questions usually have different answers.*

For example, if one treats the set M as the object of interest, and if one seeks to cover the true set M with 95% probability with some set $\hat{M}$ of μ, then $\hat{M}$ will typically be larger than the true M. Conversely, in my own paper Uhlig (2005b), I was interested in μ as the object of interest. I imposed a Bayesian prior on μ. Consequently, I was interested in finding a Bayesian confidence set $\hat{M}$, so that the true μ lies in that set with 95% probability. This set will typically be smaller than the true M. The two versions of $\hat{M}$ differ, because they answer different questions.

The same principle applies to the issue of using a classical versus a Bayesian perspective. I personally prefer the Bayesian perspective, for a number of reasons. I find it conceptually easier to focus attention entirely on potentially true μs, for which the sign restrictions surely hold, rather than having to think about confidence interval of μs, that covers a given truth with some probability. What should be in that confidence interval? Are μs allowed, that result in violations of the sign restrictions? The classical perspective requires a bit more careful thought, is more challenging and feels less natural. But once again, it is up to the researcher to decide on the question, and it is up to the econometricians to develop the appropriate tools for answering it. If someone prefers a question that requires the classical perspective, I shall not be the one to judge. Just be aware that different questions or different approaches naturally generate different answers.

Consider first what happens asymptotically. Principle 5 formulated that sign restrictions and the assumption of independent shocks deliver a one-dimensional set of θs or μs. Graphically, this is shown in Figure 9. The left panel shows, how the θs are translated into μs, using the example of no correlation between prices and quantities. For that example, the resulting set M of μs will be $M = [\pi/2, \pi]$. The right panel then shows the posterior over μs, given an infinite sample and given the sign restrictions and given a flat prior of μs on $[0, 2\pi)$. The prior and the posterior obviously coincide on that set

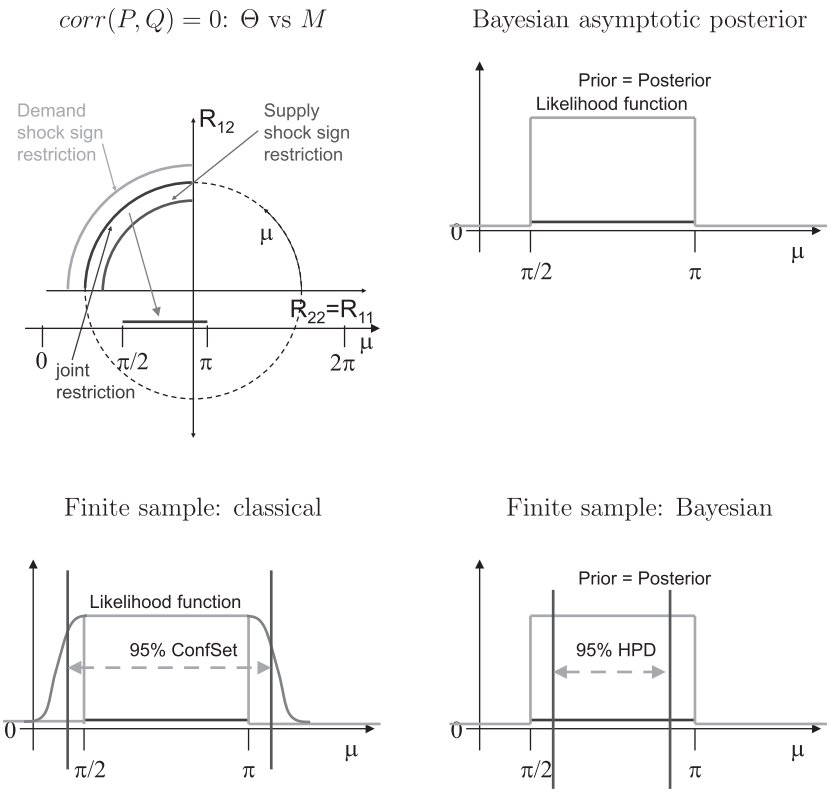

Figure 9 Inference for μ: Classical vs Bayesian, asymptotic vs finite sample

M, up to scale. This is a fairly direct consequence of Principle 5 and there is nothing really particularly surprising or remarkable about that. Let me remark on it nonetheless, as a principle.

Principle 9: *Asymptotically, the likelihood function is flat and the prior and the posterior coincide with each other, up to scale, on the set Θ of θs or the set M of μs, satisfying the sign restrictions.*

While fairly obvious, this particular point received considerable emphasis in Baumeister and Hamilton (2015).

As a Bayesian, one can now construct a set $\hat{M}$, so that the true μ lies in that set with 95% probability. For example,

$$\hat{M} = [(1 + 0.025) * \pi/2, (2 - 0.025) * \pi/2] \tag{19}$$

will do the trick in the example here. It is these sorts of sets that I have constructed in Uhlig (2005b).

Bizzarly, Fry and Pagan (2011) seem to criticize that approach, writing on p. 948 that *"... referring to this range as if it is a confidence interval ... is quite false ... it should not be imbued with probabilistic language."* This statement of theirs is either plain wrong or just confused, as a statement about the Bayesian approach advocated in Uhlig (2005b). Their statement may perhaps be true from a classical perspective. But for a Bayesian, that set $\hat{M}$ obviously has a probabilistic interpretation, as I have just explained. For those that have bothered to read their paper: there really is no "multiple shocks" or "multiple models" issue, given the Bayesian approach. It seems to me that Fry and Pagan (2011) changed the question and then elevated the changed answer to a critique of the former, ignoring Principle 8. I may have misread their paper, though. For example, perhaps they did not mean to address the approach in Uhlig (2005b) at all.

Moon and Schorfheide (2012) is an excellent investigation into the difference between classical and Bayesian inference. They note that these two approaches differ even asymptotically. A Bayesian 95% highest-posterior-density (HPD) set such as $\hat{M}$ of equation (19) is "strictly inside" the true M, whereas a classical 95% confidence set is "strictly outside." One can see that in the bottom row of Figure 9, which compares the finite-sample Bayesian and classical perspective. Both the classical econometrician and the Bayesian econometrician seem to answer the same question, namely "what is the smallest interval $\hat{M}$, so that the true μ is in $\hat{M}$ with 95% probability?" The classical econometrician holds the true μ fixed in this exercise, and is concerned with the random $\hat{M}$. The classical econometricians will then note that the true μ could be any value in the true M. The estimation confidence interval $\hat{M}$ then needs to contain the true μ with 95% probability, no matter which particular μ is the truth. For example, if the true μ happens to be equal to $\pi/2$ in the example at hand, the set $\hat{M}$ needs to start to the left of that, to be sure to include $\pi/2$. Given that the true μ is not known, the classical confidence set needs to be wider than M. The Bayesian econometrician, on the other hand, treats the sample and therefore the calculated $\hat{M}$ as given, and instead treats μ as random. If $\hat{M}$ starts to the right of $\pi/2$, it does indeed exclude the potentially true value $\mu = \pi/2$. But that's alright, since it is unlikely that the random μ is very near the boundaries of the interval $[\pi/2, \pi]$, given a uniform prior over $[0, 2\pi)$. I find it remarkable that classical and Bayesian answers often agree very closely, in seeming violation of Principle 8. Cases where the answers differ and where Principle 8 then is affirmed, as in Moon and Schorfheide (2012) or in Sims and Uhlig (1991), therefore receive our deserved attention. Just as an aside, because the correlation between price and quantity is now uncertain, given a finite sample, the likelihood function is no longer flat and the choice for the Bayesian HPD is no longer arbitrary, in contrast to what one would learn from thinking about equation 19. It also means that the bottom row of Figure 9 is slightly misleading in that regard. But this is a finite sample artifact that disappears asymptotically.

3 SHOCK IDENTIFICATION IN BVARs

Ever since Sims (1980), vector autoregressions or VARs have become a core tool in empirical macroeconomics, and are often estimated with Bayesian methods or as BVARs. There are good reasons to use them for macroeconomic data. Dynamics and lags matter; observations are rarely *i.i.d.* Moreover, one is often interested in the interrelationship of more than two data series. With that, and hopefully with Uhlig (2005b), sign restrictions have become particularly popular in the analysis of BVARs in recent years.

In reduced form, a VAR can be written as

$$Y_t = c + \sum_{j=1}^{k} B_j Y_{t-j} + u_t, \tag{20}$$

$$\text{where} \quad 0 = E[u_t \mid (Y_{t-j})_{j>0}] \text{ and } \Sigma = E[u_t u_t' \mid (Y_{t-j})_{j>0}]$$

and where Y_t is the column data vector at time t, c is a column vector of constants, B_j are square coefficient matrices, u_t is the column vector of one-step-ahead prediction errors and Σ is its variance–covariance matrix. All these can be estimated using data on Y_t. What one then often cares about, though, is to analyze the response to a structural shock, such as a monetary policy shock or a technology shock. Thus, let ϵ_t be the column vector of structural shocks. Assume them to be independent of each other, following Principle 3, and normalize their variance to unity:

$$0 = E[\epsilon_t \mid (Y_{t-j})_{j>0}] \text{ and } I = E[\epsilon_t \epsilon_t' \mid (Y_{t-j})_{j>0}]. \tag{21}$$

One then needs to state, or find, how these structural shocks map into one-step-ahead prediction errors (leaving aside the thorny issue of fundamentalness). There should then be a matrix A with

$$u_t = A\epsilon_t. \tag{22}$$

Therefore,

$$\Sigma = AA'. \tag{23}$$

Typically, it is assumed that A is square, so that there are exactly as many structural shocks as observed time series. This may be a bit odd. If another time series is added, another structural shock is added as well: why should that be the case? Certainly and as per Principle 4, it is wise to have at least as many structural shocks as time series, to avoid singularities in Σ. One could, though, conceive of having more shocks than observable series; see, e.g., Schmitt-Grohe and Uribe (2012) for a fully structural macroeconomic model of that sort. Nonetheless, we shall stick with this common practice of having equally many structural shocks as observable time series, and a square matrix A for the purpose of the exposition here.

As is well known, one can stack all this into a huge VAR of lag one. Let

$$\tilde{Y}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-k+1} \end{bmatrix}, \tilde{c} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tilde{B} = \begin{bmatrix} B_1 & \cdots & B_{k-1} & B_k \\ I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \end{bmatrix}, \tilde{A} = \begin{bmatrix} A \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

One then has

$$\tilde{Y}_t = \tilde{c} + \tilde{B}\tilde{Y}_{t-1} + \tilde{A}\epsilon_t.$$

Moreover, the impulse response at horizon $k = 0, 1, \ldots$ to the j-th shock, one standard deviation in size and of positive sign is given by

$$\tilde{r}_k = \tilde{B}^k \tilde{A} e_j$$

where e_j is a vector of the same length as ϵ with only zero entries, except a 1 in entry j. From $\tilde{r}_k$, it is now easy to extract the impulse responses of the individual time series. Sign restrictions are then typically imposed on entries in $\tilde{r}_k$ at various horizons k.

As a benchmark example, consider the focus question of understanding the effect of a monetary policy shock. It may be sensible to impose the condition that a contractionary monetary supply shock raises interest rates and lowers the money in circulation, as well as prices, for some time following the initial shock. Of course, in this day and age and given the constraints of the zero lower bound, it may no longer be wise to employ the linear specification shown above. However, there is a large literature, which has used this framework in the past, and it is very useful for discussing the issues at hand. It is an interesting research challenge to consider how to then update it all in this new age of monetary policy.

I shall consider a VAR with $k = 5$ lags in the four monthly time series of real GDP, CPI, FFR (i.e., the Federal Funds Rate) and M1, from 1959–2015. I downloaded the data from the website of the Federal Reserve Bank in St Louis. For monthly real GDP, I used a rescaled version of industrial production, and may call it industrial production in some plots or explanations. I used logs of all variables, except the FFR. I did not difference the data. In my calculations, I also dispensed with calculating standard errors or posterior analysis, which one surely should not do for a serious analysis. Obviously, something more sophisticated could be done, and is routinely done, but this shall do for this discussion: my main aim here is to clarify some issues, not to somehow provide a well-crafted econometric analysis of the focus question.

Figure 10 shows the resulting one-step-ahead prediction errors from the VAR, for FFR and industrial production. That figure may look oddly familiar. Indeed, I have used that same figure as Figure 1 to illustrate the discussion of identifying supply and demand shocks. Indeed, the same issues there arise now here, with the additional challenges created by having four rather than two dimensions and with having to consider dynamic issues.

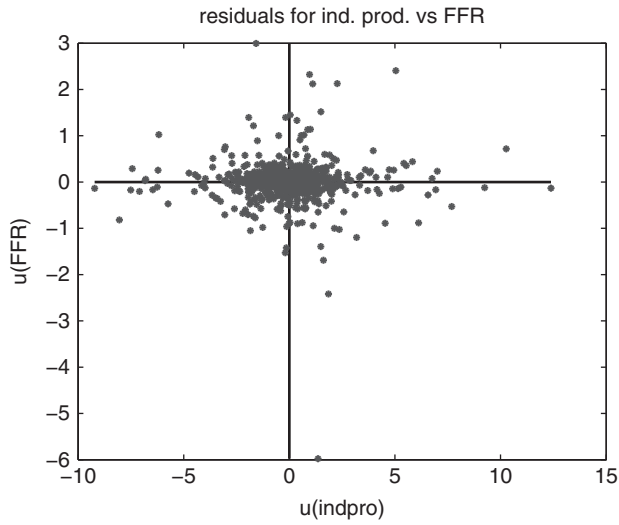

Figure 10 One-step-ahead prediction errors from the VAR, for FFR and industrial production

The identification problem here then is to disentangle the equilibrium observations of the one-step-ahead prediction errors for (realY,CPI,FFR,M1) into monetary policy shocks and other shocks. The raw data alone is insufficient. Additional identifying restrictions are needed. The key issue at hand is: which restrictions should one impose?

One popular procedure is a Cholesky decomposition. For that and the purpose of identifying monetary policy shocks, many researchers follow the lead of Bernanke and Mihov (1998) and distinguish between "slow-moving" variables that do not respond to monetary policy shocks within the period, and "fast-moving" variables that do. A Cholesky decomposition, where one orders the slow-moving variables first in any order, then the Federal Funds Rate as the monetary policy variable, and then all the fast-moving variables in any order, will provide the sought answer to the question of the response to a monetary policy shock, if that shock is identified as the Cholesky decomposition shock for the Federal Funds Rate. Among the "slow-moving" variables, one often picks production and prices on the grounds that, say, prices are sticky, while the "fast-moving" variables often include financial variables such as money stocks, other interest rates or stock market prices, given that they often react within minutes to economic news.

There is considerable appeal in that approach. Furthermore, the Cholesky decomposition is convenient, clear, and easy to compute. However, issues arise that should lead one to feel uncomfortable with that approach. One set of issues are conceptual. The other set of issues are the empirical implications.

For the first, consider the "slow-moving" variables; say, prices. Prices may be rather sticky indeed. If they were completely sticky, then they would never

move and their one-step-ahead prediction error would be zero: obviously, then, they could not react to any news at all, including news about monetary policy. However, the one-step-ahead prediction errors for prices are decidedly not zero. So even if many prices are sticky, clearly some prices react within the period to some news. By what logic, then, can these prices react to such news, but not to monetary policy shocks? The popular slate of New Keynesian models, for example, has all firms that just have been visited by the "Calvo fairy" react to all contemporaneous shocks, including monetary policy shocks. Models that impose the timing assumption inspired by Bernanke and Mihov (1998) and routinely used in empirical work can be written down, but they often quickly feel rather artificial. In sum, then, it seems to me that the logic of "slow-moving" variables not reacting to monetary policy shocks is on very thin grounds. Let me state this as follows.

Principle 10: *The Cholesky decomposition is convenient, clear. However, the "slow–fast" logic is hard to defend.*

For the second, empirical set of issues, one should note that results from Cholesky decomposition can be at considerable odds with prior views on what, say, a monetary policy shock does. Consider Figure 11, which shows the result for the VAR at hand, admittedly without any error bands. A contractionary shock in monetary policy leads to a sustained rise in prices, before prices drop: in the figure above, they actually never drop and even the resulting impulse response for inflation, i.e., the first difference of (log) CPI does not drop below zero either. It is hard to square that behavior with the conventional idea that a surprise tightening in monetary policy should lead to a throttling, not an acceleration of inflation.

There seem to be a number of responses to such results. One is to embrace them, and to call for a modified paradigm in which, indeed, contractionary monetary policy shocks raise rather than lower prices and inflation. One can write down theories in which this is the case: in some ways, it is even all too easy to come up with such theories. Some feel happy with proceeding this way. But most do not. The second possibility, then, is to declare the positive responses to be there, but not significant, and make statements using standard errors and so forth. From a Bayesian perspective, this makes little sense. Each impulse response shown by a Bayesian econometrician is a candidate truth, and not some estimator with some distance from the possible truth. If a Bayesian econometrician shows an impulse response exceeding zero, that means that some probability is assigned to this being the truth. Someone who prefers the classical perspective can wriggle out of this conundrum by citing standard errors and such, but it seems to me that this is a cop-out, rather than addressing the vexing challenge at hand. A third response is to keep trying various time series or orderings, until the problem goes away. But such a specification search should really then be done on a more formal level, as Leamer (1978) has proposed. The reasons for disregarding one specification and embracing

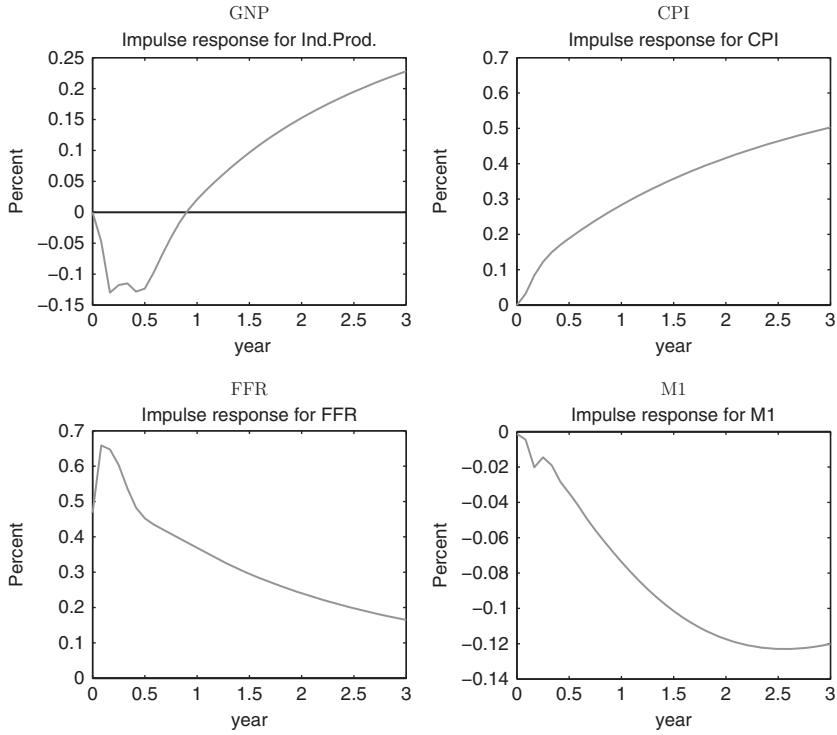

Figure 11 Cholesky decomposition results: the price puzzle

another should be spelled out for others to judge and see, and be an explicit part of the procedure, even if one does not wish to be fully formal about it. If one proceeds this way, one no longer truly uses just a Cholesky decomposition to identify shocks. Conversely, if one does, the first response listed above seems to me to be the only intellectually honest one, but not one that many might be prepared to live with. Let me formulate this as a principle:

Principle 11: *If you are worried by the "price puzzle" and the like, do not rely on the Cholesky decomposition. Use it if you are willing to "live or die" by its implications.*

"Live or die" can be dangerous. Results change over time. The price puzzle seems to have gotten worse over time. It may be wiser to abandon the Cholesky decomposition as identification procedure than to stick with it through thick and thin.

There are various other approaches, of course, some of them based entirely on the time series at hand, some based on additional information like data at a higher frequency. This is not the place to review this large and productive literature, except perhaps for pointing out that I am even more skeptical about long-run restrictions than Cholesky decomposition, for a variety of reasons;

see, e.g., Uhlig (2004). For an excellent discussion and review of much recent VAR literature, the reader is instead pointed to Ramey (2016).

3.1 Sign Restrictions in VARs

Let me turn to sign restrictions instead. I have used them for analyzing monetary policy shocks in Uhlig (2005b). Here, I wish to broaden that discussion and raise further issues.

For example, in my original article, Uhlig (2005b), I only restricted the response to a monetary policy shock. Should one sign-restrict other shocks as well? We have seen in the demand–supply discussion, above, that they can potentially help a lot. They require, though, that additional shocks can be named and their impact on the variables at hand can be signed. Recall the first two Principles 1 and 2: if you know it, impose it; otherwise do not! The same holds true for additional restrictions that one may wish to impose, such as contemporaneous causal ordering or structural relationship between some variables or some long-run restrictions like long-run monetary neutrality.

But if one only feels comfortable with sign restrictions, and if one is interested in the impact of monetary policy on output in particular, without wishing to prejudge the outcome, and if only sign restrictions for monetary policy shocks are available, then this is the list:

$$\begin{bmatrix} realY \\ CPI \\ FFR \\ M1 \end{bmatrix} = \begin{array}{c} \begin{matrix} \epsilon_m & \epsilon_1 & \epsilon_2 & \epsilon_3 \end{matrix} \\ \begin{bmatrix} ? & ? & ? & ? \\ - & ? & ? & ? \\ + & ? & ? & ? \\ - & ? & ? & ? \end{bmatrix} \end{array} \tag{24}$$

This is what I pursued in Uhlig (2005b), using six rather than four variables in the VAR. I showed that not much can be said about the reaction of output, as a result: output may react positively or negatively. If one truly believes output to react negatively, then this must arise from some additional identifying restrictions, other data, or a priori reasoning.

An important caveat with the sign restrictions (24) is that one should be reasonably sure not only that no other shock has the same sign implications as the monetary policy shock, but that this is also true for linear combinations of other shocks. The latter actually has always struck me as a potentially rather strong exclusion restriction, and one that should merit a much deeper and more critical investigation than is available in the literature so far. One example of doing so is Wolf (2016).

Note that equation (24) imposes sign restrictions on the matrix A. Some recent discussions in the literature have taken up the issue as to whether one ought to impose sign restrictions on the matrix A^{-1} instead, or perhaps even on both. The matrix A encodes the instantaneous reaction of variables to shocks. A^{-1} instead encodes such things as a monetary policy rule in reaction to observable data. For example, the equation for the Federal Funds Rate would

then relate the one-step-ahead prediction errors $u_{FFR,t}$ to $u_{GDP,t}$, $u_{CPI,t}$, $u_{M1,t}$ and the monetary policy shock $\epsilon_{m,t}$ as

$$u_{FFR,t} = \alpha u_{GDP,t} + \beta u_{CPI,t} + \gamma u_{M1,t} + \delta \epsilon_{m,t} \tag{25}$$

where α, β, γ and δ can be calculated from the appropriate coefficients in A^{-1} and Σ. In principle, one can then proceed to impose sign restrictions or even zero restrictions on some or all of these coefficients. One important recent paper arguing in favor of that approach is Arias, Caldara, and Rubio-Ramírez (2015). They argue that the Federal Reserve Bank will not react contemporaneously to the one-step-ahead prediction error in total reserves or non-borrowed reserves, perhaps on grounds that these are under the control of the Fed, and they argue for imposing sign restrictions on α.

While I applaud their hard work and serious investigation of the issues, and while I applaud the much-needed search for additional restrictions, I find that approach, and thus their results, dubious and unconvincing for the following reasons. Equation (22) implies that, potentially, all shocks move the one-step-ahead prediction errors for all equations. It may be tricky to truly label these other shocks. The Federal Reserve Bank is tasked with keeping inflation stable. One reading of the reaction function of the Fed as encapsulated in equation (25), then, is that it needs to aim at somewhat neutralizing shocks that otherwise would have subsequently lead to higher inflation down the road. That line of reasoning can potentially lead to interesting additional sign restrictions: one would have to find out what moves inflation, aside from monetary policy shocks, and then impose a reaction function to these shocks that move inflation in the opposite direction. Let me encourage interested readers to pursue this idea. However, it just seems implausible to me that these restrictions will lead to simple and direct sign restrictions or zero restrictions on A^{-1}. For example, for imposing a positive sign on α, one would have to argue that all upticks in current GDP lead to inflation down the road, if they are not caused by monetary policy. To me, that seems too strong an assumption.

An alternative reading of imposing signs or zero restrictions on the coefficients in equation (25) is that this is somehow based on what one knows how the Fed "reacts" to current news. I just feel that the Fed is doing a much more sophisticated signal-extraction exercise than would be assumed by such a simplistic rule. Moreover, one needs to realize that the one-step-ahead prediction errors appearing on the right-hand side of (25) are already moved by the monetary policy shock itself. At the very least, one would like to think of the Fed as reacting to that portion in these one-step-ahead prediction errors, that are not caused by its own actions. Finally, I cannot help to think that (25) is bringing the broken Cholesky decomposition in through the backdoor. In a Cholesky decomposition, the FFR would indeed react to all the one-step-ahead prediction errors of the "slow-moving" variables, and not react to the one-step-ahead prediction errors of the "fast-moving" variables. We have seen where that reasoning led us, see Principles 10 and 11. Should we really go down that route once again?

For the same reasons, I disagree with the statements in Baumeister and Hamilton (2015), that our economic intuition is about restricting A^{-1}. In my own thinking, I find it much more fruitful to reason from shocks to propagation and outcomes. The Fed may face the non-enviable task of partially inverting that mapping, in order to sort out which movements in observables indicate future inflationary pressure. It is certainly a valuable perspective to consider a policy maker who has to interpret observable data rather than the unobservable shocks to reach policy conclusions: in the end, the mapping surely is from observable data to policy. One may be able to circumvent reasoning about shocks that way, in the end, and I do not want to be misunderstood as discouraging thinking along on those lines: on the contrary. But once again, I find it a genuine challenge to derive straightforward restrictions about A^{-1} from that perspective.

Let me formulate this as a principle, with the caveat that there are now other researchers who disagree with it. It is good to keep this a topic of debate.

Principle 12: *Impose sign restrictions on A, not on A^{-1}, unless excellent reasons are given.*

Principle 1 and 2 are at work: I can imagine circumstances where one would rather impose restrictions on A^{-1} or on a combination of both. This is why I included the caveat in the statement of Principle 12.

3.2 Inference in BVARs

It may be good to discuss likelihood functions, prior and posteriors in BVARs. With (20) and the additional assumption, that u_t are normally distributed, one can explictly write out the likelihood function for the parameter vector ϕ of the reduced-form VAR

$$\phi = (c, B_1, \ldots, B_k, \Sigma).$$

Conditional on the initial observation, one can show that the likelihood function is proportional to a Normal-Wishart density, which specifies that

- Σ^{-1} is Wishart,
- and, conditionally on Σ, the vector of coefficients $\mathbf{b} = \text{vec}(B_1, \ldots, B_k)$ follows a Normal distribution $\mathcal{N}(\bar{\mathbf{b}}, \Sigma \otimes \mathbf{N}^{-1})$, for some $\mathbf{N}$.

This insight, which is only seemingly at odds with much of the literature on unit roots and cointegration, was at the heart of Sims and Uhlig (1991) when juxtaposing classical and Bayesian inference, and it is further elaborated upon in Uhlig (1994b). Let me formulate this as a principle.

Principle 13: *With normally distributed one-step-ahead prediction errors, the likelihood function, conditional on the initial observations, is proportional to a Normal-Wishart density.*

From this follows another interesting principle for Bayesian analysis, which has received nearly no attention so far, to the best of my knowledge.

Principle 14: *The special Normal-Wishart shape of the conditional likelihood function considerably restricts the space to where the posterior can be taken.*

The posterior is the product of the prior times the Normal-Wishart conditional likelihood function. In particular, if the prior has itself this Normal-Wishart shape, then this remains true for the posterior as well. The likelihood function has a very particular and restricted shape, though, restricting the directions in which the prior can be moved to result in the posterior. One can see that from the statement above: there are Kronecker products, implying certain proportionalities in many pairings of variables. These implicit restrictions on where the posterior can be moved, relative to the prior, and, therefore, these restrictions as to which aspects the data can speak about and which aspects it cannot, is a fascinating topic all on its own, that truly deserves much more attention than it has received so far in the literature. Let me hope that this remark spawns some insightful research on this topic in the future.

The likelihood function is Normal-Wishart, but that is true only conditionally, on the initial observations. The initial observations can be thought of as containing considerable information about the VAR parameters, if they are viewed as drawn from the same data generating process. One can most easily see this for an AR(1) with a fixed variance for the innovation term. If the root for the AR(1) is much closer to zero than to one in absolute value, the unconditional distribution of a draw from this time series will be reasonably tight around zero. If the root is much closer to unity, an unconditional draw will tend to be drawn from a density that is spread out far more. The issues and the resulting implications for Bayesian priors and Bayesian analysis are discussed in more detail in Uhlig (1994a) and Uhlig (1994b).

As a matter of practice, one may often proceed with the following procedure to generate inference for a BVAR with sign restrictions:

1. Take many draws $\mathbf{b}_i$, Σ_i.
2. Cholesky-decompose $\Sigma_i = L_i L_i'$.
3. Any other decomposition $\Sigma_i = A_i A_i'$ satisfies

$$A_i = L_i R_i, \text{ where } R_i R_i' = I.$$

4. Draw R_i, see below. Check whether A_i satisfies sign restrictions.
5. Check sign restrictions.

If the restriction is on a single shock, then it suffices to obtain a single column of A_i. For that, it then suffices to draw a single vector r_i with $|| r_i ||= 1$ ("first column of R_i") and calculate $a_i = L_i r_i$ ("first column of A").

That procedure should be clear enough, except for this portion: how to draw R_i and from which distribution? As far as the distribution is concerned, there is one that is particularly natural. Recall that any other decomposition $\Sigma_i = A_i A_i'$ satisfies

$$A_i = L_i R_i, \text{ where } R_i R_i' = I$$

It is desirable to have a procedure so that the particular choice of the decomposition $\Sigma_i = L_i L_i'$ does not matter. For example, the original ordering of the variables should result in the same probability distribution for the resulting A_i or a_i as any other ordering: otherwise the whole issue of ordering is brought back in through the back door. But that means, it should be equally likely to draw R_i or $\tilde{R}_i = QR_i$, where $QQ' = I$. This is called the Haar measure, as is well understood and as has been pointed out particularly clearly by Rubio-Ramírez, Waggoner, and Zha (2010). With one vector only, it should be equally likely to draw r_i or $\tilde{r}_i = Qq_i$, where $QQ' = I$. Put differently, the distribution of r_i should be uniform on a sphere. This, at least, is easy to accomplish: draw each entry from a standard normal distribution and normalize the resulting vector to unit length.

For some reason, Baumeister and Hamilton (2015) criticize this Haar measure, since transformations of a uniform priors look informative. Of course they do. That does not invalidate the principle advantage of using a Haar measure, in my view.

Principle 15: *Use the Haar measure to draw R_i or r_i.*

A subtle issue arises in the sampling procedure as stated above, which concerns the interaction of sign restrictions and the reduced-form VAR. The sampling procedure described samples from a joint posterior for $(\mathbf{b}, \Sigma, R)$. Consider, then, two different $(\mathbf{b}, \Sigma)$. If the sign restrictions are easy to satisfy for the first, but difficult to satisfy for the second, the first will be sampled more often, i.e., be given higher marginal probability relative to the no-sign-restriction case. But "difficult to satisfy" means the sign restrictions are better at tightly identifying the shock; see Principle 7. Thus, the procedure above implicitly puts more weight on $(\mathbf{b}, \Sigma)$, which offer less tight identification. It may be better to use a conditional prior instead:

1. Pick some standard prior for the reduced form $(\mathbf{b}, \Sigma)$.
2. Take a draw $(\mathbf{b}, \Sigma)$ from the resulting posterior.
3. Check whether there is any R then satisfying sign restrictions.
4. If not, discard.
5. If yes, draw a fixed number that satisfy sign restrictions.

While perhaps more appealing, it seems harder to implement in practice, since step 3 should rely on analytics, not random draw testing. In any case, a proper foundation per a Bayesian prior for either sampling procedure can be given and is provided in Uhlig (2005b), if that helps to inform the choice.

When imposing sign restrictions, the issue arises for how many periods after the shock they shall be imposed. In my discussion Uhlig (1998), I showed that imposing it for longer than just the initial impact can make a considerable difference. Theory can potentially provide some guidances. Thus, Canova and Paustian (2011) provide a model without capital and where doubts then arise about the persistence of the signs of the shock response, calling sign restrictions beyond the initial period into questions. Note also that the price–quantity framework discussed in the first half of this chapter has no persistence of shocks: I considered an *i.i.d.* example on purpose. In fact, it is easy to construct non-persistent theories. But are these convincing reasons to exclude extended horizons, when imposing sign restrictions? Put differently, are non-persistent theories plausible from the perspective of a priori reasoning? Now, if you truly trust and believe some non-persistent theory, or can convince an audience within reason that it offers the right framework, as opposed to one with persistence, do go ahead; see Principle 1. But do not, if you do not; see Principle 2. Usually, many macroeconomic theories are highly stylized on purpose: one should probably feel uncomfortable trusting all their implications. I therefore conclude that it is typically appropriate to impose sign restrictions beyond the initial impact period.

Principle 16: *Often, it is reasonable to impose sign restrictions for up to one year after the initial shock.*

It is instructive to analyze, how longer-horizon sign restrictions make a difference in simple time series models. Consider, for example, a bivariate VAR with one lag. Suppose that, for some reason, the coefficient matrix is always diagonalizable, and has two identical roots and stable roots. Let me distinguish three cases, then. One case is that the two roots are real and positive. In that case, the sign of the initial shock persists forever, i.e., an impulse response to a positive shock stays positive. The second case is that the two roots are real and negative. In that case, the impulse response will oscillate between negative and positive values, period by period. The third case is that the two roots are complex conjugate. In that case, we get oscillations that may last longer than a single period: think "damped sin waves." Consider then imposing the sign restriction that the response is positive, up to some horizon K, where $K = 0$ would mean imposing sign restrictions on impact only. These sign restrictions do not further inform case 1. They rule out case 2, in case $K > 0$. Further, they rule out ranges of complex roots, and the more this is so, the larger is K: only damped sin waves with the lowest frequencies will show a positive response for sufficiently many periods after the impact. While I shall leave the

details of the algebra to the reader, the insight from this example should be clear: dynamic sign restrictions can have considerable bite. They do not have to: indeed, if it is known that the roots are real and positive, increasing K has no bite at all.

Finally, let me offer some remarks on how to read impulse response ranges. With a Bayesian analysis, these are proper probability statements about posteriors. The common way of plotting them and their quantiles provides graphic information about marginal distributions, at each horizon and for each variable plotted. Each single plot therefore provides information about a number of marginal distributions. The median of the distribution of some impulse response at some horizon is the median of that particular marginal distribution. It is obvious that connecting the medians from these marginals does not represent some particular draw or even some "median" draw. They simply show how the medians and these marginal distributions evolve, from one horizon to the next. This has always been true for Bayesian impulse response error bands, and I find it rather unremarkable. I would not have remarked on it if it weren't for Fry and Pagan (2011) on page 949 somehow thinking this to be a big deal. Personally, I am much less inclined to highlight this as a key insight, but nevertheless:

Principle 17: *The common way of plotting Bayesian impulse responses and their quantiles provides graphic information about marginal distributions, at each horizon and for each variable plotted. The median of the distribution of some impulse response at some horizon is the median of that particular marginal distribution. As a collection, they typically do not represent some particular draw or even some "median" draw.*

4 CONCLUSIONS AND SOME FINAL RECOMMENDATIONS

Rather than summarizing the chapter (which I have already done, in essence, in the introduction), let me conclude by offering some final thoughts and recommendations. Sometimes sign restrictions are criticized because they are weak and sometimes not much can be concluded. They may be weak indeed. But, truly, what else is there? Principles 1 and 2 hold: if more can comfortably be imposed, go ahead, but if sign restrictions are all one has, then one has no choice but to live with the weak conclusions that can be drawn on their basis, or change the research strategy entirely. For example, while I find Cholesky decompositions easy and clear to use, and while they sometimes can be put to excellent and informative purpose, their defense is sometimes murky and their results unpalatable; see Principles 10 and 11. Long-run restrictions are often even more problematic, in my view, and seem to require exact knowledge of unit roots and such. Perhaps medium-run restrictions are a way out of some of the conundrums bedeviling long-run restrictions – see, e.g., Uhlig (2004) – but it would be a stretch to say that they have found much popularity so far.

There are lots of other approaches that are promising and which sometimes truly deliver insightful answers. I am thinking of approaches such as variance-maximization procedures, exploiting regional data or panel data, employing high-frequency identification approaches, introducing regime shifts or heteroskedasticity induced by shocks around, say, FOMC announcement dates (with all due apologies to the key authors proposing these approaches, for not providing a more fully fledged and citation-based review of that literature). When a debate such as the one laid out in this chapter can make some approach thorny, one is then tempted to think that the grass is greener over there. But, as Friedman has remarked, there is no such thing as a free lunch. It is generally useful advice to avoid losing the general lessons learned from VARs on aggregate time series when turning to these other approaches. On the contrary: they need to be thought through anew for these approaches as well. And with that, and with the discussion offered here, the sign restriction approach should be even more appealing.

I have not much left now, therefore, but to once again mention my main recommendation. Recall **Principles 1 and 2:** *if you know it, impose it! If you do not know it, do not impose it!*

References

Angrist, J. D., K. Graddy, and G. W. Imbens (2000). "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *The Review of Economic Studies 67*, 499–527.

Arias, J. E., D. Caldara, and J. F. Rubio-Ramírez (2015). "The Systematic Component of Monetary Policy in Svars: An Agnostic Identification Procedure." *International Finance Discussion Papers, no 1131*.

Baumeister, C. and J. D. Hamilton (2015, September). "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information." *Econometrica 83*(5), 1963–1999.

Bernanke, B. S. and I. Mihov (1998, August). "Measuring Monetary Policy." *The Quarterly Journal of Economics 113*(3), 869–902.

Canay, I. and A. M. Shaikh (2017). "Practical and Theoretical Advances for Inference in Partially Identified Models." In *Advances in Economics and Econometrics (Proceedings of the 11th World Congress of the Econometric Society)*. Cambridge, UK: Cambridge University Press Chapter 9, pp. 271–306.

Canova, F. and G. de Nicolo (2002). "Monetary Disturbances Matter for Business Fluctuations in the G-7." *Journal of Monetary Economics 49*(6), 1131–1159.

Canova, F. and M. Paustian (2011). "Business Cycle Measurement With Some Theory." *Journal of Monetary Economics 58*, 345–361.

Canova, F. and J. P. Pina (1999). "Monetary Policy Misspecification in Var Models." CEPR Discussion Papers 2333, Center for Economic and Policy Research.

Danne, C. (2015, December). "Varsignr: Estimating Vars Using Sign Restrictions in R." *Munich Personal RePEc Archive, MPRA Paper No. 68429*.

Dwyer, M. (1998). "Impulse Response Priors for Discriminating Structural Vector Autoregressions." Working paper 780, UCLA.

Faust, J. (1998). "The Robustness of Identified Var Conclusions About Money." *Carnegie-Rochester Conference Series on Public Policy 49*, 207–244.

Fry, R. and A. Pagan (2011). "Sign Restrictions in Structural Vector Autoregressions: A Critical Review." *Journal of Economic Literature 49*(4), 938–960.

Ho, K. and A. M. Rosen (2017). "Partial Identification in Applied Research: Benefits and Challenges." In *Advances in Economics and Econometrics (Proceedings of the 11th World Congress of the Econometric Society)*. Cambridge, UK: Cambridge University Press Chapter 10, pp. 307–359.

Kline, B. and E. Tamer (2016). "Default Bayesian Inference in a Class of Partially Identified Models." Draft, Harvard University.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference With Nonexperimental Data.* Wiley series in probability and mathematical statistics. New York: John Wiley and Sons, Inc.

Leamer, E. E. (1981, August). "Is It A Demand Curve, or Is It A Supply Curve? Partial Identification Through Inequality Constraints." *The Review of Economics and Statistics 63*(3), 319–327.

McCloskey, D. N. (1998). *The Rhetoric of Economics* (2nd ed.). Madison: University of Wisconsin Press.

Moon, H. R. and F. Schorfheide (2012, March). "Bayesian and Frequentist Inference in Partially Identified Models." *Econometrica 80*(2), 755–782.

Ramey, V. (2016). "Macroeconomic Shocks and Their Propagation." In J. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2 of *Handbook of Macroeconomics*. Elsevier Chapter 2, pp. 71–162.

Rubio-Ramírez, J. F., D. F. Waggoner, and T. Zha (2010). "Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference." *Review of Economic Studies 77*(2), 665–696.

Schmitt-Grohe, S. and M. Uribe (2012, November). "What's News in Business Cycles." *Econometrica 80*(6), 2733–2764.

Sims, C. A. (1980). "Macroeconomics and Reality." *Econometrica 48*, 1–48.

Sims, C. A. and H. Uhlig (1991). "Understanding Unit Rooters: A Helicopter Tour." *Econometrica 59*, 1591–1599.

Uhlig, H. (1994a). "On Jeffrey's Prior When Using the Exact Likelihood Function." *Econometric Theory 10*(3–4), 633–644.

Uhlig, H. (1994b). "What Macroeconomists Should Know About Unit Roots: A Bayesian Perspective." *Econometric Theory 10*(3–4), 645–671.

Uhlig, H. (1998). "The Robustness of Identified Var Conclusions About Money: A Comment." *Carnegie-Rochester Conference Series on Public Policy 49*, 245–263.

Uhlig, H. (2004, April–May). "Do Technology Shocks Lead to a Fall in Total Hours Worked?" *Journal of the European Economic Association 2*(2–3), 361–371.

Uhlig, H. (2005a). "Supply, Demand and Identification." Draft, Humboldt Universität zu Berlin.

Uhlig, H. (2005b). "What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure." *Journal of Monetary Economics 52*, 381–419.

Wolf, C. K. (2016). "A Characterization of Identified Impulse Response Sets in Invertible Linear State-Space Models." Draft, Princeton University.