9 Multivariate & Nonlinear Time Series Analysis

The previous chapter discussed univariate processes, whereby a single variable y_t is measured over time. We speak of a multivariate process if y_t is vector-valued, e.g. $y_t = (y_{1t}, y_{2t})' \in \mathbb{R}^2$. After some introductory remarks on multivariate processes in Section 9.1, Section 9.2 discusses Vector Autoregressions (VARs), arguably the most popular tool in empirical macroeconomics. Building on the insights obtained from VARs, Section 9.3 analyzes state space models. As it turns out, most if not all time series models can be written in state space form. This includes multivariate as well as univariate processes like ARMA(p, q) models, and it includes linear as well as nonlinear models, going even so far to encompass models for which the exact representation for y_t (or its conditional expectation) as a function of past values might not even be available in closed form.

9.1 Multivariate Processes

Stationarity and ergodicity of a multivariate process are defined analogously to the univariate case. In particular, weak stationarity requires the mean $\mu_t = \mathbb{E}[y_t]$ and autocovariances $\Gamma_{h,t} = \mathbb{E}[(y_t - \mu_t)(y_{t-h} - \mu_{t-h})']$ to be constant over time. For a multivariate process $y_t \in \mathbb{R}^n$, μ_t is an $n \times 1$ vector containing the means of the individual series in y_t, $\mathbb{E}[y_{it}]$ for $i = 1 : n$, and $\Gamma_{h,t}$ is an $n \times n$ matrix. For example, in the bivariate case ($n = 2$), we have

$$\Gamma_{h,t} = \begin{bmatrix} \text{Cov}(y_{1,t}, y_{1,t-h}) & \text{Cov}(y_{1,t}, y_{2,t-h}) \\ \text{Cov}(y_{1,t-h}, y_{2,t}) & \text{Cov}(y_{2,t-h}, y_{2,t-h}) \end{bmatrix} \equiv \begin{bmatrix} \gamma_{11,h} & \gamma_{12,h} \\ \gamma_{21,h} & \gamma_{22,h} \end{bmatrix} .$$

The diagonal elements are simply the univariate autocovariances of y_{1t} and y_{2t}. The off-diagonal element $\gamma_{12,h}$ is the cross-covariance between y_{1t} and y_{2t} with the former leading the latter by a displacement of h lags. Note that $\gamma_{12,h} \neq \gamma_{21,h}$ because the correlation of y_{1t}

with past movements in y_{2t} is (generally) not the same as the correlation of y_{2t} with past movements in y_{1t}. As a result, in contrast to the univariate case, $\Gamma_h \neq \Gamma_{-h}$. Instead, we have $\gamma_{12,h} = \gamma_{21,-h}$ and therefore $\Gamma_h = \Gamma'_{-h}$.[1]

A multivariate White Noise (WN) process and the multivariate General Linear Process (GLP) are also defined analogously to the univariate case. A mean-zero WN process $u_t \sim WN(0, \Sigma)$ is defined by

$$\mathbb{E}[u_t] = 0 \; , \quad \mathbb{E}[u_t u'_{t-h}] = \begin{cases} \Sigma & \text{if } h = 0 \\ 0 & \text{otherwise} \end{cases} \; .$$

Note that Σ is not necessarily diagonal, i.e. the individual components of the vector-valued WN process u_t can be contemporaneously correlated. However, the defining feature of a WN process is that each of the components is uncorrelated with its own past (and future) movements as well as those of any of the other components. The multivariate GLP is defined as

$$y_t = B(L)u_t = \sum_{l=0}^{\infty} B_l u_{t-l} \; , \quad \text{with } u_t \sim WN(0, \Sigma) \; , \quad B_0 = I \; , \quad \sum_{l=0}^{\infty} ||B_l||^2 < \infty \; .$$

This leads to the ACF $\Gamma_h = \mathbb{E}\left[\left(\sum_{l=0}^{\infty} B_l u_{t-l}\right)\left(\sum_{k=0}^{\infty} B_k u'_{t-h-k}\right)\right] = \sum_{k=0}^{\infty} B_{k+h} \Sigma B'_k$ for $h \geq 0$.

9.2 Vector Autoregressions (VARs)

A VAR is a common way to approximate the multivariate GLP. A VAR(p) models the series y_t as a linear function of the past p lags of y_t:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u_t \; , \quad u_t \sim WN(0, \Sigma) \; , \tag{9.1}$$

where Φ_0 is an $n \times 1$ vector and $\{\Phi_l\}_{l=1:p}$ are $n \times n$ matrices. For example, for a bivariate VAR(1) with mean zero ($\Phi_0 = 0$), we have

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \; , \quad \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \sim WN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \; , \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right) \; .$$

There are two sources of interaction between y_{1t} and y_{2t}. First, the value of each series today depends on its own, but also on the other variable's value yesterday. Second, the innovations

[1] $\Gamma'_{t,-h} = \mathbb{E}[(y_t - \mu_t)(y_{t+h} - \mu_{t+h})']' = \mathbb{E}[(y_{t+h} - \mu_{t+h})(y_t - \mu_t)'] = \Gamma_{t+h,h}$ and under WS, the timing does not matter: $\Gamma_{t+h,h} = \Gamma_{t,h} = \Gamma_h$.

to the two series are potentially correlated (contemporaneously). In the general VAR(p), we can write the equation for each series y_{it} as

$$\begin{aligned} y_{it} = \Phi_{c,1} &+ \Phi_{1,i1}y_{1,t-1} + ... + \Phi_{1,in}y_{n,t-1} \\ &+ \Phi_{2,i1}y_{1,t-2} + ... + \Phi_{2,in}y_{n,t-2} \\ &+ \quad ... \\ &+ \Phi_{p,i1}y_{1,t-p} + ... + \Phi_{p,in}y_{n,t-p} + u_{it} \ , \end{aligned}$$

i.e. each variable y_{it} depends on its own p lags, p lags of each of the other series in y_t, as well as its own innovation u_{it}. The latter may be correlated with the innovations to the other variables in y_t.

Stationarity As already discussed in Section 8.2, a VAR(1) is weakly stationary (WS) iff all eigenvalues of Φ_1 are below one in absolute value. To analyze stationarity of the VAR(p), we write it in companion form as a (restricted) VAR(1):

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ \\ y_{t-p+1} \end{bmatrix}}_{x_t} = \underbrace{\begin{bmatrix} \Phi_0 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}}_{F_0} + \underbrace{\begin{bmatrix} \Phi_1 & \Phi_2 & ... & & \Phi_p \\ I & 0 & ... & & 0 \\ 0 & I & & & \vdots \\ \vdots & & \ddots & & \\ 0 & & & I & 0 \end{bmatrix}}_{F_{np \times np}} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ \\ y_{t-p} \end{bmatrix}}_{x_{t-1}} + \underbrace{\begin{bmatrix} u_t \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}}_{v_t} .$$

The VAR(p) is WS iff this companion form-VAR(1) is WS, i.e. iff all eigenvalues of F are below one in absolute value.[2]

Moments & GLP-Representation Once we know that a VAR(p) is WS, it is quite trivial to calculate its mean. For example, for a VAR(1), $\mathbb{E}[y_t] = \mu \ \forall \ t$ implies $\mu = \Phi_0 + \Phi_1\mu$, which yields $\mu = (I - \Phi_1)^{-1}\Phi_0$. Analogously, under p lags, we get $\mu = (I - \Phi_1 - ... - \Phi_p)^{-1}\Phi_0$.

To obtain the autocovariances $\{\Gamma_h\}_{h=0,1,...}$, it is useful to write the VAR(p) in GLP-form. Just as the AR(p) process can be written in the form of the univariate GLP, i.e. as an MA(∞), so too we can write the VAR(p) in the form of the multivariate GLP, i.e. as a

[2] Note the analogy to the stationarity discussion for univariate AR(p) processes in Section 8.2. It explained that an AR(1) is weakly stationary (WS) if the autoregressive parameter ϕ_1 is smaller than one in absolute value, whereas, to analyze stationarity of an AR(p), we write it in companion form as a restricted VAR(1), and we demand that the autoregressive matrix of parameters F has all eigenvalues below one in absolute value.

VMA(∞). Consider first a VAR(1). Repeatedly inserting for lags y_{t-l}, $l = 1, 2, ...$, we get

$$y_t = \sum_{l=0}^{\infty} \Phi_1^l \Phi_0 + \sum_{l=0}^{\infty} \Phi_1^l u_{t-l} + \lim_{k \to \infty} \Phi_1^k y_{t-k} \ .$$

If Φ_1 has all eigenvalues in the unit circle (and the process was initialized at a finite value in the very distant past), $\lim_{k\to\infty} \Phi_1^k y_{t-k} = 0$. Based on this expression, we get the ACF

$$\Gamma_h = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)'] = \mathbb{E}\left[\left(\sum_{l=0}^{\infty} \Phi_1^l u_{t-l} \right) \left(\sum_{k=0}^{\infty} \Phi_1^k u_{t-h-k} \right)' \right] = \sum_{k=0}^{\infty} \Phi_1^k \Sigma \Phi_1^{k+h\prime} \ ,$$

for $h \geq 0$. Because $\lim_{k\to\infty} \Phi_1^k = 0$, we can approximate Γ_h well by cutting off the infinite sum after, say, K terms: $\Gamma_h \approx \sum_{k=0}^{K} \Phi_1^{k+h} \Sigma \Phi_1^{k\prime}$ for K large.[3] [4]

To write a VAR(p) in GLP-form, we first write its companion-form-VAR(1) – $x_t = F_0 + F x_{t-1} + v_t$ – in GLP-form:

$$x_t = \sum_{l=0}^{\infty} F^l F_0 + \sum_{l=0}^{\infty} F^l v_{t-l} \ .$$

Given x_t, we can get y_t as $y_t = M x_t$, where $M = [I_n, 0_n, ..., 0_n]$ is an $n \times (np)$ matrix that selects the first n elements of the $np \times 1$ vector x_t. As a result, we obtain

$$\begin{aligned} y_t &= \sum_{l=0}^{\infty} M F^l F_0 + \sum_{l=0}^{\infty} M F^l v_{t-l} \\ &= \sum_{l=0}^{\infty} (F^l)_{11} \Phi_0 + \sum_{l=0}^{\infty} (F^l)_{11} u_{t-l} \ , \end{aligned}$$

where $(F^l)_{11}$ denotes the upper-left $n \times n$ block of the $np \times np$ matrix F^l. The second equality follows by the structure of M, F_0 and v_t. From this, we obtain

$$\Gamma_h^y = \sum_{k=0}^{\infty} (F^k)_{11} \Sigma (F^{k+h})'_{11} \ .$$[5]

[3] It is also possible to get the ACF of a VAR(1) exactly. First, use the fact that $\mathbb{V}[y_t] = \mathbb{V}[y_{t-1}]$ under stationarity to write

$$\Gamma_0 = \mathbb{V}[y_t] = \mathbb{V}[\Phi_1 y_{t-1} + u_t] = \Phi_1 \Gamma_0 \Phi_1' + \Sigma \ .$$

Using the rule that $vec(ABC) = (C' \otimes A) vec(B)$, we get that $vec(\Gamma_0) = (\Phi_1 \otimes \Phi_1) vec(\Gamma_0) + vec(\Sigma) = [I - (\Phi_1 \otimes \Phi_1)]^{-1} vec(\Sigma)$. Given Γ_0, we obtain Γ_h for $h \geq 1$ using the Yule-Walker first-order difference equation for the ACF of a VAR(1): $\Gamma_h = \Phi_1 \Gamma_{h-1}$ for $h \geq 1$.

[4] Note that the GLP-form of the VAR(p) also shows us that $\mu = \sum_{l=0}^{\infty} \Phi_1^l \Phi_0$. This equals $(I - \Phi_1)^{-1} \Phi_0$ because $\sum_{l=0}^{\infty} \Phi_1^l = (I - \Phi_1)^{-1}$.

[5] Again, an exact expression for Γ_h^y can be obtained by first obtaining the exact Γ_h^x for the VAR(1) x_t as

Reduced-Form vs. Structural Representation The above expressions for the VAR(p) are referred to as its reduced-form representation, and the WN-terms u_t that appear in it are called reduced-form errors. They are the forecasting errors obtained when predicting y_t one step ahead, i.e. using information available at time $t-1$, $\mathscr{F}_{t-1} = \{y_{t-l}\}_{l=1}^{\infty}$:

$$u_t = y_t - \mathbb{E}_{t-1}[y_t] = y_t - \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} \ ,$$

where $\mathbb{E}_{t-1}[y_t]$ is shorthand for $\mathbb{E}[y_t|\mathscr{F}_{t-1}]$. This is why u_t is also referred to as the innovation to the process y_t at time t; it contains everything that affects y_t and is not known to the researcher at time $t-1$. As mentioned above, u_t can be cross-sectionally (contemporaneously) correlated, i.e. any u_{it} can be correlated with any u_{jt}. Therefore, the derivative

$$\frac{\partial y_{i,t+h}}{\partial u_{j,t}} = \frac{\partial y_{i,t+h}}{\partial y_{j,t}} = \left[(F^h)_{11}\right]_{ij}$$

is only useful from a predictive point of view.[6] It tells us how useful y_{jt} is in predicting the series y_{it} h periods into the future. This predictive notion of causality is referred to as Granger- or Granger-Sims-causality. Because u_t is cross-sectionally correlated, we do not know whether a change in u_{jt} (and hence y_{jt}) indeed induces a change in $y_{i,t+h}$ or whether there is a third series $y_{m,t}$, correlated with y_{jt} (as u_{mt} is correlated with u_{jt}), that causes a change in $y_{i,t+h}$. A famous quote, attributed to John Cochrane, says that "the weather forecast Granger-causes the weather, but shooting the weatherman will not produce a sunny weekend."

To make causal statements in a VAR requires us to decomopose the reduced-form errors into the underlying, independent driving forces of y_t, referred to as shocks. We can write

$$u_t = \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ , \tag{9.2}$$

where ε_t is the vector of shocks. Note that $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ has to hold. ε_t is often assumed to be

described in the footnote above. Then, we can compute

$$\Gamma_h^y = \mathbb{E}[(y_t - \mu)(y_{t-h} - \mu)'] = \mathbb{E}[(M(x_t - \mu^x))(M(x_{t-h} - \mu^x))'] = M\Gamma_h^x M' = (\Gamma_h^x)_{11} \ .$$

Also analogously as under a VAR(1), the GLP-form here shows us that $\mu = \sum_{l=0}^{\infty}(F^l)_{11}$ and, therefore, that $(I - \Phi_1 - ... - \Phi_p)^{-1} = \sum_{l=0}^{\infty}(F^l)_{11}$.

[6] Starting from the equation $y_{t+h} = \Phi_0 + \Phi_1 y_{t+h-1} + u_{t+h}$ and repeatedly inserting for y_{t+h-l}, $l = 1, 2, ...,$, it is easy to see that the coefficient in front of y_t is the same as that in front of u_t. The same holds more generally for a VAR(p), but the calculation goes via the companion form $x_{t+h} = F_0 + Fx_{t+h-1} + v_{t+h}$.

strict WN, i.e. the individual shocks ε_{jt} are not only uncorrelated, but fully independent. Also, it is usually taken to be n-dimensional, just as u_t, i.e. there are n independent shocks in ε_t driving the n innovations in u_t. This leads to the structural representation of the VAR(p):

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ . \tag{9.3}$$

Sometimes, it is also written as

$$Ay_t = B_0 + B_1 y_{t-1} + ... + B_p y_{t-p} + \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ , \tag{9.4}$$

where there is a one-to-one mapping between the two expressions, with $A = \Phi_\varepsilon^{-1}$ and $B_l = \Phi_\varepsilon^{-1}\Phi_l$ for $l = 0, 1, ..., p$. The approach of thinking about causality in a system of equations with endogenous variables by thinking of shocks as their underlying driving forces was pioneered by Chris Sims, Nobel prize winner in 2011.

Suppose that we know all parameters in the VAR, including the matrix Φ_ε that maps reduced-form errors into structural shocks. This allows us to compute impulse-response functions (IRFs), variance decompositions (VDs) and historical decompositions (HDs).

An IRF illustrates the dynamic effects of a shock on a series in y_t. Concretely, the effect of shock j, ε_{jt}, on series i, y_{it}, h periods into the future is given by the derivative

$$\frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}} = \left[(F^h)_{11}\Phi_\varepsilon\right]_{ij} \ . \tag{9.5}$$

The sequence $\left\{\frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}}\right\}_{h=0,1,...}$ is referred to as the IRF of variable i to shock j. It tells us the dynamic effect of ε_{jt} on the series y_{it}. Because, by definition, ε_{jt} is i.i.d., this effect has a causal interpretation.

A VD of y_t determines the contribution of different shocks to the variance (and autocovariances) of y_t. Recall that the variance of y_t is given by

$$\Gamma_0 = \sum_{l=0}^{\infty}(F^l)_{11}\Sigma(F^l)'_{11} = \sum_{l=0}^{\infty}(F^l)_{11}\Phi_\varepsilon I \Phi'_\varepsilon (F^l)'_{11} \ .$$

Let $I^{(j)}$ be the identity matrix with all but the jth diagonal entry set to zero. Then

$$\Gamma_0^{(j)} = \sum_{l=0}^{\infty}(F^l)_{11}\Phi_\varepsilon I^{(j)} \Phi'_\varepsilon (F^l)'_{11}$$

is the part of the variance of y_t that is due to the variation of shock ε_{jt}, whereby the matrix

$I^{(j)}$ allows us to set the variances of all shocks but shock j to zero. If we set the variances of all shocks to zero, the variance of y_t would be zero as well. Also, by virtue of Γ_0 being a linear function in the diagonal elements of $I^{(j)}$, we have that $\Gamma_0 = \sum_{j=1}^n \Gamma_0^{(j)}$. As a result, the fraction of the variance of a particular series y_{it} – $\mathbb{V}[y_{it}] = \left[\Gamma_0\right]_{ii}$ – that is explained by ε_{jt} is given by

$$\left[\Gamma_0^{(j)}\right]_{ii} / \left[\Gamma_0\right]_{ii} \; . \tag{9.6}$$

Analogous ratios can also be computed for contemporaneous covariances $\text{Cov}(y_{it}, y_{jt})$. Moreover, by computing the contribution of a shock to the autocovariance Γ_h for some $h \geq 1$, we can compute such ratios also for autocoviarance terms of the form $\text{Cov}(y_{it}, y_{j,t-h})$.

Finally, an HD of $\{y_t\}_{t=1}^T$ determines the contribution of different shocks to the actually observed evolution of y_t in the data. Given initial conditions $\{y_t\}_{t=-p+1:0}$, the data $\{y_t\}_{t=1}^T$ is determined by the shocks $\{\varepsilon_t\}_{t=1}^T$ via the set of VAR-equations:

$$y_t = \Phi_0 + \sum_{l=1}^p \Phi_l y_{t-p} + \Phi_\varepsilon I \varepsilon_t \; , \quad t = 1:T \; .$$

If we set all shocks to zero at all periods, then the series y_t would be determined only by its initial conditions, whose effect vanishes in the long term under stationarity, which means that y_t gradually converges to its unconditional mean. Setting all but the jth shock to zero, we get a hypothetical series $\{y_t^{(j)}\}_{t=1}^T$ that would have been obtained were it only for shock j:

$$y_t^{(j)} = \Phi_0 + \sum_{l=1}^p \Phi_l y_{t-p}^{(j)} + \Phi_\varepsilon I^{(j)} \varepsilon_t \; , \quad t = 1:T \; , \tag{9.7}$$

where $y_t^{(j)} = y_t$ for the initial conditions $t = -p+1:0$. Of course, in any practical application, we do not know the "true" past shocks $\{\varepsilon_t\}_{t=1}^T$ (just as we do not know $\Phi_0, \{\Phi_l\}_{l=1:p}$ and Φ_ε), but we estimate them using $\hat{\Phi}_\varepsilon$ and the estimated reduced-form errors $\{\hat{u}_t\}_{t=1}^T$. Based on $\{y_t^{(j)}\}_{t=1}^T$, we can compute various statistics like $\hat{\mathbb{V}}[y_{it}^{(j)}]/\hat{\mathbb{V}}[y_{it}]$, the historical contribution of shock j to the variance of y_{it}, or we can determine the impact of shock j on the value of series i at a particular date t.

9.2.1 Estimation of Reduced-Form VARs

We can write the VAR(p),

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u_t \ , \quad u_t \sim WN(0, \Sigma) \ ,$$

in linear regression form as follows. Let $k = np + 1$ and define

$$\underset{(T\times n)}{Y} = [y_1, ..., y_T]' \ , \quad \underset{(T\times n)}{U} = [u_1, ..., u_T]' \ ,$$

as well as

$$\underset{(k\times 1)}{x_t} = [y_{t-1}', ..., y_{t-p}', 1]' \ , \quad \underset{(T\times k)}{X} = [x_1, ..., x_T]' \ , \quad \text{and} \quad \underset{(k\times n)}{\Phi} = [\Phi_1, ..., \Phi_p, \Phi_0]' \ .$$

Then

$$y_t' = x_t'\Phi + u_t' \ , \quad \text{and} \quad Y = X\Phi + U \ .$$

Frequentist Inference Under normality of u_t, we have the conditional density

$$\begin{aligned} p(y_t|Y_{t-p:t-1}, \Phi, \Sigma) &= (2\pi)^{-n/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(y_t' - x_t'\Phi)\Sigma^{-1}(y_t' - x_t'\Phi)'\right\} \\ &= (2\pi)^{-n/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(y_t' - x_t'\Phi)'(y_t' - x_t'\Phi)\right]\right\} \ , \end{aligned}$$

where the second line uses the fact that $a'Ba = tr[Baa']$. We get the conditional likelihood

$$\begin{aligned} p(Y_{1:T}|Y_{-p+1:0}, \Phi, \Sigma) &= \prod_{t=1}^{T} p(y_t|Y_{t-p:t-1}, \Phi, \Sigma) \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\sum_{t=1}^{T}\frac{1}{2}tr\left[\Sigma^{-1}(y_t' - x_t'\Phi)'(y_t' - x_t'\Phi)\right]\right\} \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\sum_{t=1}^{T}\Sigma^{-1}(y_t' - x_t'\Phi)'(y_t' - x_t'\Phi)\right]\right\} \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right]\right\} \ , \end{aligned}$$

using the facts that $tr[A+B] = tr[A]+tr[B]$ and $\sum_{t=1}^{T}(y_t' - x_t'\Phi)'(y_t' - x_t'\Phi) = (Y - X\Phi)'(Y - X\Phi)$. Maximizing this expression by taking derivatives gives the ML estimators

$$\hat{\Phi} = (X'X)^{-1}X'Y \ , \quad \hat{\Sigma} = \frac{1}{T}(Y - X\hat{\Phi})'(Y - X\hat{\Phi}) \ .^{7}$$

The same $\hat{\Phi}$ is obtained under Least Squares (LS):

$$\hat{\Phi} = \underset{\Phi}{\arg\min} \sum_{t=1}^{T} (y_t' - x_t'\Phi)\Sigma^{-1}(y_t' - x_t'\Phi)' = \underset{\Phi}{\arg\min}\ tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right] .$$

Note that $\hat{\Phi}$ is composed of equation-by-equation ML/LS estimators: $\hat{\Phi} = [\hat{\Phi}^{(1)}, ..., \hat{\Phi}^{(n)}]$, where $\hat{\Phi}^{(i)} = (X'X)^{-1}X'Y^{(i)}$ is the ML/LS estimator in the regression $y_{it} = x_t'\Phi^{(i)} + u_{it}$, and where $Y^{(i)} = Y_{\cdot i} = [y_{i1}, ...y_{iT}]'$ is the ith column of Y, containing only data for series y_{it}. Therefore, to obtain $\hat{\Phi}$, we can also run n separate regressions for each row in the VAR(p).

The insights from Section 8.4 go through. First, the asymptotic properties of $\hat{\Phi}$ can be analyzed using the LLN and CLT for ergodic and strictly stationary (SS) time series. Second, if y_t has a unit-root, the asymptotic distribution of $\hat{\Phi}$ is not Normal, and even for roots close to but below unity, its finite sample distribution is far from Normal. Furthermore, following the discussion from Section 6.2, to compute standard errors and construct confidence intervals for functions of Φ and Σ, like Γ_h^y, we need an analytical expression for the (finite sample) distribution of $\hat{\Gamma}_h^y$ as a function of the true Γ_h^y. We can approximate it by the corresponding asymptotic distribution, obtained using the Delta method. However, this approximation is poor if y_t is close to a unit-root process.

Bayesian Inference For Bayesian analysis of a VAR, a popular class of prior distributions for (Φ, Σ) is the (Matrix-)Normal-Inverse Wishart (MNIW) prior:

$$\Phi|\Sigma \sim MN(\underline{\mu}, \underline{P}^{-1}, \Sigma)\ , \quad \Sigma \sim IW(\underline{S}, \underline{\nu})\ .^{8}$$

This is the analogue to the (Multivariate-)Normal-Inverse Gamma prior for (β, σ^2) in the linear regression model, which notably includes the AR(p) model, the univariate version of the VAR(p) (see Section 4.5 and Section 8.4). As shown in the Appendix, following the same steps as in Section 4.5, we get the MNIW posterior:

$$\Phi|Y, \Sigma \sim MN(\bar{\mu}, \bar{P}^{-1}, \Sigma)\ , \quad \Sigma|Y \sim IW(\bar{S}, \bar{\nu})\ ,$$

[7] The derivations use the rules $tr[A] = tr[A']$, $tr[A+B] = tr[A] + tr[B]$ as well as

$$\frac{\partial tr[AXB]}{\partial X} = A'B'\ , \quad \frac{\partial tr[X'BXC]}{\partial X} = BXC + B'XC'\ , \quad \frac{\partial tr[AX^{-1}B]}{\partial X} = -(X^{-1}BAX^{-1})'\ , \quad \frac{\partial ln|X|}{\partial X} = (X')^{-1}\ .$$

[8] As discussed in Appendix B, the former is equivalently stated as $vec(\Phi)|\Sigma \sim N(\underline{\mu}, \Sigma \otimes \underline{P}^{-1})$.

with

$$\bar{P} = \underline{P} + X'X \ , \qquad \bar{\mu} = \bar{P}^{-1}[X'Y + \underline{P}\underline{\mu}] \ ,$$

$$\bar{\nu} = \underline{\nu} + T \ , \qquad \bar{S} = \underline{S} + Y'Y + \underline{\mu}'\underline{P}\underline{\mu} - \bar{\mu}'\bar{P}\bar{\mu} \ .$$

Taking the improper prior $p(\Phi, \Sigma) \propto c$ instead, we get the same expression, but with $\bar{P} = X'X$, $\bar{\mu} = \hat{\Phi}$, $\bar{S} = (Y - X\hat{\Phi})'(Y - X\hat{\Phi})$ and $\bar{\nu} = T - k - n - 1$.[9] Given the posterior for (Φ, Σ), we can easily compute (numerically) the posterior of any function of them, like Γ_h^y, based on which we can construct credible sets (see Section 6.2).

How to specify a prior for the parameters (Φ, Σ) in a model as complex as the VAR? First, note that sometimes priors are used simply as a regularization technqique. For example, one can estimate the VAR by shrinking coefficients in Φ to zero using a Lasso or Ridge prior (see Section 4.5). Instead, actual initial beliefs on the dynamics of y_t can be shaped into a prior for (Φ, Σ) by using dummy observations Y^* and X^*. These dummy observations could be actual data from another country, observations generated by simulating a macroeconomic model, or they could be generated by introspection, as under the Minnesota prior (see Appendix), which postulates a list of properties believed to hold in the data (e.g. that a series tends to persist at its current level, i.e. is a unit root). Based on dummy observations, we can form, for example, the MNIW prior

$$\Phi|\Sigma \sim MN(\underline{\mu}, \Sigma \otimes \underline{P}^{-1}) \ , \quad \Sigma \sim IW(\underline{S}, \underline{\nu}) \ ,$$

with $\underline{P} = X^{*\prime}X^*$, $\underline{\mu} = \hat{\Phi}^*$, $\underline{S} = (Y^* - X^*\hat{\Phi}^*)'(Y^* - X^*\hat{\Phi}^*)$ and $\underline{\nu} = T^* - k - n - 1$, where T^* is the dimension of Y^* and X^*. This is the posterior that would be obtained if our dummy observations (Y^*, X^*) were the true data and if we used an improper prior. As a result, this prior reflects the view that (Φ, Σ) should be such that they generate observations Y^* and X^*.

[9] The derivation applies the same steps as under the MNIW-prior above. Thereby, $p(\Phi, \Sigma) \propto c$ implies $p(\Phi|\Sigma) \propto c$ and $p(\Sigma) \propto c$. Another commonly used prior is $p(\Phi, \Sigma) \propto |\Sigma|^{(n+1)/2}$, which implies $p(\Phi|\Sigma) \propto c$ and $p(\Sigma) \propto |\Sigma|^{(n+1)/2}$ and leads to the MNIW-posterior with $\bar{P} = X'X$, $\bar{\mu} = \hat{\Phi}$, $\bar{S} = (Y - X\hat{\Phi})'(Y - X\hat{\Phi})$ and $\bar{\nu} = T$.

9.2.2 Estimation of Structural VARs

Above, we referred to the structural representation of the VAR:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ ,$$

where $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ because $u_t = \Phi_\varepsilon \varepsilon_t \sim N(0, \Sigma)$. The previous section showed how to estimate (Φ, Σ). This section is devoted to finding Φ_ε, which in turn allows us to analyze causal relationships between shocks ε_t and endogenous variables y_t by computing IRFs, VDs and HDs.

It turns out that Φ_ε is not identified (from the data), only $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ is. This is because only $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ – not Φ_ε itself – appears in the likelihood associated with the Structural VAR (SVAR) equation above:

$$\begin{aligned} p(Y_{1:T}|Y_{-p+1:0}, \Phi, \Phi_\varepsilon) &= \prod_{t=1}^{T} p(y_t|Y_{t-p:t-1}, \Phi, \Phi_\varepsilon) \\ &= (2\pi)^{-nT/2} |\Phi_\varepsilon \Phi_\varepsilon'|^{-T/2} exp \left\{ - \sum_{t=1}^{T} \frac{1}{2} (y_t' - x_t'\Phi) \left(\Phi_\varepsilon \Phi_\varepsilon'\right)^{-1} (y_t' - x_t'\Phi)' \right\} \\ &= (2\pi)^{-nT/2} |\Sigma|^{-T/2} exp \left\{ - \sum_{t=1}^{T} \frac{1}{2} (y_t' - x_t'\Phi) \Sigma^{-1} (y_t' - x_t'\Phi)' \right\} \ . \end{aligned}$$

There are $n(n+1)/2$ unique elements in the symmetric matrix Σ, while there are n^2 elements in Φ_ε. This means that we cannot pinpoint Φ_ε uniquely based on data, not in finite samples and neither asymptotically. Instead, there is a set of values for (the elements in) Φ_ε that are in line with the data. We say that Φ_ε is not point-identified, but set-identified. The problem of reducing this identified set is referred to as shock identification.

To point-identify Φ_ε, we need (at least) $n^2 - n(n+1)/2 = n(n-1)/2$ (point) restrictions/identification assumptions. Even with fewer restrictions, we can decrease the identified set, and we can possibly point-identify the effects of some of the n shocks in ε_t (by point-identifying the corresponding columns of Φ_ε) or the responses of some of the n series in y_t (by point-identifying the corresponding rows of Φ_ε).

We can further analyze and eventually address this identification problem by decomposing Φ_ε into a part that we can point-identify and a part that we cannot point-identify. Let Ω be any orthogonal matrix, i.e. it holds that $\Omega\Omega' = I$. Note that we can always do a Cholesky decomposition of Σ to get the unique lower-triangular matrix Σ_{tr}, which is such that $\Sigma = \Sigma_{tr}\Sigma_{tr}'$. Based on these properties of Ω and Σ_{tr} as well as the fact that $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$, we can write Φ_ε as $\Phi_\varepsilon = \Sigma_{tr}\Omega$. Because we can point-identify Σ and because the Cholesky decomposition is unique, we can uniquely identify Σ_{tr} from the data. In contrast, any Ω is consistent with the data (i.e. gives the same likelihood), because for any Ω, we have $\Phi_\varepsilon\Phi_\varepsilon' = \Sigma_{tr}\Omega\Omega'\Sigma_{tr}' = \Sigma$. From a frequentist point of view, this means that we cannot find a (unique) point estimator for Ω from the data. From a Bayesian point of view, it means that any prior on Ω is not updated by the data:

$$\begin{aligned} p(\Phi, \Sigma, \Omega|Y) &\propto p(Y|\Phi, \Sigma, \Omega)p(\Phi, \Sigma, \Omega) \\ &= p(Y|\Phi, \Sigma)p(\Phi, \Sigma, \Omega) \\ &= p(Y|\Phi, \Sigma)p(\Phi, \Sigma)p(\Omega|\Phi, \Sigma, \cdot) \\ &\propto p(\Phi, \Sigma|Y)p(\Omega|\Phi, \Sigma, \cdot) \ . \end{aligned} \tag{9.8}$$

Because the likelihood is independent of Ω, the joint posterior of (Φ, Σ, Ω) is proportional to the (marginal) posterior of (Φ, Σ) – which is unaffected by Ω – times our prior for Ω.[10] [11] Any conclusions we draw for Ω come from our prior for it!

This decomposition of Φ_ε and analysis of the joint posterior of (Φ, Σ, Ω) is useful to understand how inference on the SVAR is conducted in practice, both Bayesian as well as frequentist. First, consider Bayesian inference. Given a prior for (Φ, Σ), we obtain the (marginal) posterior $p(\Phi, \Sigma|Y)$ as explained in Section 9.2.1 above. Based on the derivation in Eq. (9.8), a draw from the joint posterior $p(\Phi, \Sigma, \Omega|Y)$ is obtained by first drawing (Φ, Σ) from $p(\Phi, \Sigma|Y)$ and then, conditional on them, drawing Ω from the prior $p(\Omega|\Phi, \Sigma, \cdot)$. As usual under Bayesian inference, this prior reflects ex-ante beliefs of the researcher, before having seen the data. In the context of SVARs in particular, it reflects the restrictions deemed to lead to a credible identification of Φ_ε and subsequent causal analysis of the dynamcis of y_t. Examples of such restrictions are discussed below.

To understand frequentist inference for SVARs, recall that, in absence of identification issues, the ML estimator is equal to the posterior mode under a uniform prior because the posterior is proportional to the likelihood in this case. If a parameter is not identified, this posterior mode, i.e. the ML estimator, is not unique, but there is a set of values that all maximize the likelihood. Concretely, for an SVAR, any orthogonal Ω maximizes the likelihood. We can obtain the identified set of Ω – and therefore the identified set of Φ_ε and any function of it (like IRFs, VDs, etc.) – by drawing many values for Ω from a Uniform prior distribution on the space of all orthogonal matrices. We can shrink this identified set by imposing restrictions

[10] I write this prior as $p(\Omega|\Phi, \Sigma, \cdot)$ because it can and usually does depend on Φ and Σ, and sometimes even on the data Y. See the discussion of different types of restrictions below.

[11] Another way to see it: the marginal posterior for Ω is only a function of the marginal prior we assumed for it. The calculations above imply $p(\Omega|Y) = \int p(\Omega|\Phi, \Sigma, \cdot)d(\Phi, \Sigma)$.

on Ω and functions of it. Overall, frequentist inference on (Φ, Σ, Ω) is conducted by taking the point-identified $(\hat{\Phi}, \hat{\Sigma})$ and constructing the identified set for $\hat{\Omega}$ by drawing it from a flat (Uniform) prior and ensuring that the imposed restrictions are satisfied.

The bottom line is that in both the Bayesian but also the frequentist approach we need to draw from the prior $p(\Omega|\Phi, \Sigma, \cdot)$, which is such that it ensures that – conditional on Φ, Σ, and possibly the data Y – certain restrictions are satisfied. The only difference is that under Bayesian inference we draw many values of (Φ, Σ) and construct the identified set for Ω (and any function of it like IRFs, VDs, etc.) by conditioning on each of these draws, whereas under frequentist inference we consider the single, point-identified $(\hat{\Phi}, \hat{\Sigma})$ and we draw Ω from its flat prior conditioning on this particular $(\hat{\Phi}, \hat{\Sigma})$. This is summarized in Algorithms 15 and 16.

Algorithm 15 (Bayesian Inference in SVARs)**.**

```
1. Specify priors $p(\Phi, \Sigma)$ and $p(\Omega|\Phi, \Sigma, \cdot)$.
2. Using $p(\Phi, \Sigma)$, compute the marginal posterior $p(\Phi, \Sigma|Y)$.
3. For $m = 1 : M$,
    (a) draw $(\Phi^{(m)}, \Sigma^{(m)})$ from $p(\Phi, \Sigma|Y)$
    (b) draw $\Omega^{(m)}$ from $p(\Omega|\Phi^{(m)}, \Sigma^{(m)}, \cdot)$
   The set of values $\{(\Phi^{(m)}, \Sigma^{(m)}, \Omega^{(m)})\}_{m=1}^{M}$ approximates $p(\Phi, \Sigma, \Omega|Y)$ numerically.
```

Algorithm 16 (Frequentist Inference in SVARs)**.**

```
1. Specify restrictions/identification assumptions $p(\Omega|\Phi, \Sigma, \cdot)$.
2. Compute $(\hat{\Phi}, \hat{\Sigma})$.
3. For $m = 1 : M$, draw $\Omega^{(m)}$ from $p(\Omega|\hat{\Phi}, \hat{\Sigma}, \cdot)$. The set of values $\{(\hat{\Phi}, \hat{\Sigma}, \Omega^{(m)})\}_{m=1}^{M}$
   approximates the identified set of $(\Phi, \Sigma, \Omega)$ numerically.
```

Of course, the restrictions could be such that Ω is point-identified. In this case, given $(\Phi^{(m)}, \Sigma^{(m)})$ or given $(\hat{\Phi}, \hat{\Sigma})$, we take the unique value $\Omega^{(m)}$ that satisfies the restrictions. Under set-identification of Ω, computational efficiency considerations can become important if the identified set is small.[12] For more details on how to draw from $p(\Omega|\Phi, \Sigma, \cdot)$ efficiently, see Arias et al. (2018).

[12] With sign restrictions, we could in principle draw $\Omega^{(m)}$ from an unrestricted (Uniform) distribution and then throw away any draws that do not satisfy the sign restrictions. This can become very inefficient if the identified set is small. More efficient approaches incorporate sign restrictions directly into the prior distribution. In contrast to that, point restrictions need to be reflected in the prior directly because under a continuous distribution, the probability of obtaining a draw that exactly satisfies a point restrictions is zero.

The following paragraphs discuss possible ways of constructing $p(\Omega|\Phi, \Sigma, \cdot)$, i.e. imposing restrictions that reduce the identified set for Ω. Importantly, these methods do not exclude each other, but one might for example obtain the posterior $p(\Phi, \Sigma, \Omega|Y)$ or the identified set for (Φ, Σ, Ω) by combining point- and sign-restrictions. The following discussion focuses on restrictions imposed on Ω and Φ_ε. Equivalently, one might also impose restrictions on $A = \Phi_\varepsilon^{-1}$, using the other way to write the SVAR in Eq. (9.4).

Point-Restrictions A point-restriction fixes some of the elements of Ω. For example, one might impose a point restriction on short-term dynamics, such as real activity reacting to monetary policy shocks only with a lag. If $y_{i,t}$ is GDP and ε_{jt} is the MP shock, this implies that

$$\frac{\partial y_{i,t}}{\partial \varepsilon_{jt}} = \big((F^0)_{11}\Phi_\varepsilon\big)_{ij} = (\Sigma_{tr}\Omega)_{ij} = 0$$

(recall Eq. (9.5)) and restricts the dot product of the ith row of Σ_{tr} and the jth column of Ω to be zero.

Such short-run point-restrictions are typically based on assumptions about decision- or informational lags, which often appear dubious (Uhlig, 2017) but might be more plausible in high-frequency settings. A particular point-restriction on short-term dynamics which perfectly identifies Ω is to simply fix $\Omega = I$. In this case, $\Phi_\varepsilon = \Sigma_{tr}$, which is why this identificaton assumption is referred to as “Cholesky identification”. Because Σ_{tr} is lower-triangular, it comes equal to assuming a particular ordering of the variables in y_t such that the first variable is only affected by the first shock, the second is affected by the first two shocks, etc.

One can also point down elements in Ω by imposing point-restrictions on dynamics in the long run. For example, motivated by economic theory, Blanchard and Quah (1989) assume that the monetary policy shock does not affect GDP after 20 quarters, i.e. money neutrality holds in the long run. Under quarterly data, this assumption implies that

$$\frac{\partial y_{i,t+20}}{\partial \varepsilon_{jt}} = \big((F^{20})_{11}\Phi_\varepsilon\big)_{ij} = \big((F^{20})_{11}\Sigma_{tr}\Omega\big)_{ij} = 0 \ .$$

A downside of this approach is that, oftentimes, long-run impulse responses are usually estimated rather imprecisely, which yields imprecise estimates of Ω.

Sign-Restrictions Rather than fixing the exact values of certain impulse responses (or statistics more generally) to some values, one might impose that they have a particular sign. For example, one might assume that the monetary policy shock does not raise output upon impact. This implies that

$$\frac{\partial y_{i,t}}{\partial \varepsilon_{jt}} = (\Phi_\varepsilon)_{ij} = (\Sigma_{tr}\Omega)_{ij} \leq 0 \; .$$

Depending on Σ_{tr}, this sign restriction can shrink the identified set for Ω a lot – in the limit achieving point identification – or not at all.

Analogously, sign restrictions can also be imposed for longer-term IRFs. Moreover, we can also impose restrictions on VDs, assuming e.g. that no more than 20% of the variance in GDP is due to the monetary policy shock: $\left[\Gamma_0^{(j)}\right]_{ii} / \left[\Gamma_0\right]_{ii} < 0.1$, where both $\Gamma_0^{(j)}$ and Γ_0 are functions of $\Phi_\varepsilon = \Sigma_{tr}\Omega$.

Narrative Restrictions When sign- (or point-) restrictions are imposed on HDs, we speak of narrative restrictions. For example, we might assume that, over a particular sub-period of our sample, the monetary policy shock contributed at least 10% to the variance of GDP: $\mathbb{V}[y_{it}^{(j)}]/\mathbb{V}[y_{it}] > 0.2$, considering periods t in the defined time frame. Similarly, we might say that supply shocks contributed most to the fall in GDP at the start of the Covid-19 pandemic.

Note that under narrative restrictions, the prior for Ω is also a function of the data Y: $p(\Omega|\Phi, \Sigma, Y)$. From a Bayesian point of view, this is inocuous, as the inference conditions on Y. In contrast, from a frequentist point of view, this considerably complicates the analysis.

Identification Through Instrumental Variables One can shrink the identified set of Φ_ε also using external instruments. For example, if an IV z_t is correlated with the shock ε_{1t}, but uncorrelated with the shock ε_{2t}, this implies the following point-restriction:

$$\mathbb{E}[\varepsilon_{2t} z_t] = \mathbb{E}[I^{(2)}\varepsilon_t z_t] = \mathbb{E}[I^{(2)}\Phi_\varepsilon^{-1} u_t z_t] = \mathbb{E}[I^{(2)}(\Sigma_{tr}\Omega)^{-1} u_t z_t] = 0 \; .$$

Analogous sign-restrictions based on IVs are also imaginable. The restrictions are operationalized by replacing the expectation with the sample mean.