10 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 10.1, Section 10.2 discusses Vector Autoregressions (VARs), arguably the most popular tool in empirical macroeconomics. Building on the insights obtained from VARs, Section 10.3 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.

10.1 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 $h \geq 0$.

10.2 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:

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u_t \ , \quad u_t \sim WN(0, \Sigma) \ , \tag{10.1}$$

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.

Any VAR(p) can be written in companion form as a (restricted) VAR(1):

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ \\ y_{t-p+1} \end{bmatrix}}_{y_t^c} = \underbrace{\begin{bmatrix} \Phi_0 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}}_{F_0} + \underbrace{\begin{bmatrix} \Phi_1 & \Phi_2 & ... & & \Phi_p \\ I & 0 & ... & & 0 \\ 0 & I & & & \vdots \\ \vdots & & \ddots & & \\ 0 & & & I & 0 \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ \\ y_{t-p} \end{bmatrix}}_{y_{t-1}^c} + \underbrace{\begin{bmatrix} u_t \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix}}_{u_t^c} ,$$

where y_t^c and u_t^c are $np \times 1$-vectors, F is an $np \times np$-matrix, and $\mathbb{V}[u_t^c] = Q$ is an $np \times np$-matrix of zeros except for its upper-left $n \times n$-block, which equals Σ. Note that

$$y_t = M y_t^c \ ,$$

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 y_t^c.

Stationarity As discussed in Section 9.2, a VAR(1) is weakly stationary (WS) iff all eigenvalues of Φ_1 are below one in absolute value. A 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]

[2]Note the analogy to the stationarity discussion for univariate AR(p) processes in Section 9.2. 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.

Moments Under WS, we have $\mathbb{E}[y_t] = \mu \ \forall \ t$. For a VAR(1), this implies $\mu = \Phi_0 + \Phi_1\mu$, which yields $\mu = (I - \Phi_1)^{-1}\Phi_0$. For a VAR($p$), we can proceed analogously to get $\mu = (I - \Phi_1 - ... - \Phi_p)^{-1}\Phi_0$, or we can compute the mean of y_t via the mean of its companion form: $\mathbb{E}[y_t] = M\mathbb{E}[y_t^c]$, where $\mathbb{E}[y_t^c] = (I - F)^{-1}F_0$.

Under WS, we also have $\mathbb{V}[y_t] = \Gamma_0 \ \forall \ t$ and $\text{Cov}(y_t, y_{t-h}) = \Gamma_h \ \forall \ t$. For a VAR(1), $\mathbb{V}[y_t] = \mathbb{V}[y_{t-1}]$ implies

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

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):

$$\Gamma_h = \Phi_1\Gamma_{h-1} \ , \quad h \geq 1 \ .^3$$

For a VAR(p), we can get the autocovariances from the autocovariances of its companion form, then. If $\Gamma_h^c = \text{Cov}(y_t^c, y_{t-h}^c)$ for $h = 0, 1, ...$, then

$$\Gamma_h = M\Gamma_h^c M' = (\Gamma_h^c)_{11} \ ,$$

i.e. Γ_h equals the upper-left $n \times n$-block of Γ_h^c.

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

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

[3] Replacing Φ_0 by $(I - \Phi_1)\mu$ in $y_t = \Phi_0 + \Phi_1 y_{t-1} + u_t$ yields

$$(y_t - \mu) = \Phi_1(y_{t-1} - \mu) + u_t \ .$$

Right-multiplying by $(y_{t-h} - \mu)'$ and taking expectations yields the Yule-Walker equation.

If Φ_1 has all eigenvalues in the unit circle (and the process was initialized at a finite value in the very distant past), then $\lim_{k\to\infty} \Phi_1^k y_{t-k} = 0$.[4]

To write a VAR(p) in GLP-form, we first write its companion form-VAR(1) in GLP-form:

$$y_t^c = \sum_{l=0}^{\infty} F^l F_0 + \sum_{l=0}^{\infty} F^l u_{t-l}^c \ .$$

Using the fact that $y_t = M y_t^c$, we obtain

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

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 u_t^c.[5]

[4]This yields another expression for 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 $h \geq 0$. 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]From this, we obtain

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

Also, the GLP-form shows us that $\mu = \sum_{l=0}^{\infty} (F^l)_{11} \Phi_0$. Because $\lim_{L\to\infty} F^L = 0$, we can approximate them well by cutting off the infinite sums after some L terms:

$$\mu \approx \sum_{l=0}^{L} (F^l)_{11} \Phi_0 \ , \quad \Gamma_h \approx \sum_{l=0}^{L} \left(F^l\right)_{11} \Sigma \left(F^{l+h}\right)'_{11} \quad \text{for } L \text{ large} \ .$$

Using these expression can be computationally more efficient if the companion form is very high-dimensional due to a large p (under a large n) since no inversion of an $np \times np$-matrix is required.

Forecasting Repeatedly inserting for lags y_{t-l}, $l = 1, 2, ..., h$ in the VAR(1) equation (i.e. following analogous steps as above for the GLP/VMA(∞)-representation), we get

$$y_t = \sum_{l=0}^{h} \Phi_1^l \Phi_0 + \sum_{l=0}^{h-1} \Phi_1^l u_{t-l} + \Phi_1^h y_{t-h} \ .$$

More generally, for a VAR(p), we get

$$y_t = \sum_{l=0}^{h} (F^l)_{11} \Phi_0 + \sum_{l=0}^{h-1} (F^l)_{11} u_{t-l} + (F^h)_{1\cdot} y_{t-h}^c \ ,$$

where $(F^h)_{1\cdot}$ denotes the $n \times np$-matrix consisting of the first n rows of F^h, and $y_{t-h}^c = (y_{t-h}', y_{t-h-1}', ..., y_{t-h-p+1}')'$. This implies that the (model-consistent) h-step ahead forecast of y_t is given by

$$\hat{y}_{t|t-h} = \mathbb{E}[y_t | \mathcal{F}_{t-h}] = \sum_{l=0}^{h} (F^l)_{11} \Phi_0 + (F^h)_{1\cdot} y_{t-h}^c \ ,$$

where $\mathcal{F}_{t-h} = \{y_{t-h}, y_{t-h-1}, ...\}$ is the information set at time $t-h$. Equivalently,

$$\hat{y}_{t+h|t} = \mathbb{E}[y_{t+h} | \mathcal{F}_t] = \sum_{l=0}^{h} (F^l)_{11} \Phi_0 + (F^h)_{1\cdot} y_t^c \ .$$

Note that the same expression is obtained by iterating on the VAR equation

$$\hat{y}_{t+l|t} = \Phi_0 + \Phi_1 \hat{y}_{t+l-1|t} + ... + \Phi_p \hat{y}_{t+l-p|t} + \hat{u}_{t+l|t}$$

for $l = 1 : h$ by inserting the known expressions $\hat{y}_{t|t} = y_t$, $\hat{y}_{t-1|t} = y_{t-1}$, etc., and inserting $\hat{u}_{t+l|t} = 0$ for all $l = 1 : h$. This entails simulating the VAR for h periods starting from x_t.

Note that the (non-predictable) forecast error is given by

$$u_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t} = \sum_{l=0}^{h-1} (F^l)_{11} u_{t+h-l} \ .$$

Its variance is

$$\Psi_h \equiv \mathbb{V}[u_{t+h|t}] = \sum_{l=0}^{h-1} (F^l)_{11} \Sigma (F^l)_{11}'$$

and converges to the unconditional variance of y_t, $\mathbb{V}[y_t]$, as $h \to \infty$.

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 $t-1$, $\mathcal{F}_{t-1} = \{y_{t-l}\}_{l=1}^{\infty}$:

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

where $\mathbb{E}_{t-1}[y_t]$ is shorthand for $\mathbb{E}[y_t|\mathcal{F}_{t-1}]$. 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, we need to decomopose the reduced-form errors into the underlying independent driving forces of y_t, the shocks ε_t. We can write

$$u_t = \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ . \tag{10.2}$$

Note that $\Sigma = \Phi_\varepsilon \Phi_\varepsilon'$ has to hold. ε_t is often assumed to be 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):

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \Phi_\varepsilon \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ . \tag{10.3}$$

[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 $y^c_{t+h} = F_0 + F y^c_{t+h-1} + u^c_{t+h}$.

Sometimes, it is also written as

$$Ay_t = B_0 + B_1 y_{t-1} + ... + B_p y_{t-p} + \varepsilon_t \ , \quad \varepsilon_t \sim WN(0, I) \ , \tag{10.4}$$

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 we know both Φ and Φ_ε - the matrix that maps reduced-form errors into structural shocks. Then we can compute Impulse-Response Functions (IRFs), Variance Decompositions (VDs), Forecast Error Variance Decompositions (FEVDs) and Historical Decompositions (HDs).

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

$$\frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}} = \left[(F^h)_{11}\Phi_\varepsilon\right]_{ij} \ . \tag{10.5}$$

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 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.[7] Since Γ_0 is 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

[7] If we set the variances of all shocks to zero, the variance of y_t would be zero as well.

given by

$$\left[\Gamma_0^{(j)}\right]_{ii} / \left[\Gamma_0\right]_{ii} \ . \tag{10.6}$$

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 $\text{Cov}(y_{it}, y_{j,t-h})$.

An FEVD decomposes the variance of forecast errors $u_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$ rather than series y_t. We have

$$\Psi_h = \mathbb{V}[u_{t+h|t}] = \sum_{l=0}^{h-1} (F^l)_{11} \Sigma (F^l)'_{11} = \sum_{l=0}^{h-1} (F^l)_{11} \Phi_\varepsilon I \Phi'_\varepsilon (F^l)'_{11} \ .$$

Hence, the fraction of the variance of $u_{i,t+h|t}$ (of the mean squared error when forecasting y_{it} h periods ahead) that is due to shock j is given by

$$\left[\Psi_h^{(j)}\right]_{ii} / \left[\Psi_h\right]_{ii} \ , \quad \Psi_h^{(j)} = \sum_{l=0}^{h-1} (F^l)_{11} \Phi_\varepsilon I^{(j)} \Phi'_\varepsilon (F^l)'_{11} \ .$$

Analogous ratios can also be computed for forecasting error covariances, $\text{Cov}(u_{i,t+h|t}, u_{j,t+h|t})$.

Finally, an HD determines the contribution of different shocks to the actually observed evolution of $\{y_t\}_{t=1}^T$ 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 for all time periods, we get a hypothetical series $\{y_t^{(j)}\}_{t=1}^T$ that would have been obtained were it only for shock j:

$$y_t^{(j)} = \Phi_0 + \sum_{l=1}^{p} \Phi_l y_{t-p}^{(j)} + \Phi_\varepsilon I^{(j)} \varepsilon_t \ , \quad t = 1:T \ , \tag{10.7}$$

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.

10.2.1 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

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

10.2.1.1 Frequentist Inference

Under Least Squares (LS),

$$\begin{aligned}\hat{\Phi}_{LS} &= \arg\min_{\Phi} \sum_{t=1}^T (y_t' - x_t'\Phi)(y_t' - x_t'\Phi)' \\ &= \arg\min_{\Phi} tr\left[(Y - X\Phi)'(Y - X\Phi)\right] \\ &= (X'X)^{-1}X'Y \ .^8\end{aligned}$$

In turn, we can estimate Σ as

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^T \hat{u}_t\hat{u}_t' = \frac{1}{T}\hat{U}'\hat{U} = \frac{1}{T}(Y - X\hat{\Phi})'(Y - X\hat{\Phi}) \ .$$

[8] We use the rule that $\partial tr[Z'BZC]/\partial Z = BZC + B'ZC'$, applied to $Z = (Y - X\Phi)$ with $B = C = I$ in this case.

Note that $\hat{\Phi}$ is composed of equation-by-equation LS estimators: $\hat{\Phi} = [\hat{\Phi}^{(1)}, ..., \hat{\Phi}^{(n)}]$, where $\hat{\Phi}^{(i)} = (X'X)^{-1}X'Y^{(i)}$ is the LS estimator of $\Phi^{(i)}$ 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).

These same estimators are obtained as ML estimators under normality of u_t. In that case, 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 $\hat{\Phi}$ and $\hat{\Sigma}$ from above.[9]

The insights from Section 9.4 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 7.2, 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

[9]On top of the rule above, the derivations use $tr[A] = tr[A']$, $tr[A+B] = tr[A] + tr[B]$ as well as

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

asymptotic distribution, obtained using the Delta method. However, this approximation is poor if y_t is close to a unit-root process.

10.2.1.2 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:

$$\Phi|\Sigma \sim MN(\underline{\mu}, \underline{P}^{-1}, \Sigma) \ , \quad \Sigma \sim IW(\underline{S}, \underline{\nu}) \ .^{10}$$

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 5.1 and Section 9.4). As shown in the Appendix, following the same steps as in Section 5.1, we get the MNIW posterior:

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

with

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

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

The Marginal Data Density (MDD) is then

$$p(Y) = \pi^{-nT/2}|\underline{P}|^{n/2}|\bar{P}|^{-n/2}|\underline{S}|^{\underline{\nu}/2}|\bar{S}|^{-\bar{\nu}/2}\Gamma_n\left(\frac{\bar{\nu}}{2}\right)/\Gamma_n\left(\frac{\underline{\nu}}{2}\right) \ .$$

Instead, under the improper prior $p(\Phi, \Sigma) \propto c$, we get the analogous MNIW-posterior with

$$\bar{P} = X'X \ , \quad \bar{\mu} = \hat{\Phi} \ , \quad \bar{S} = (Y - X\hat{\Phi})'(Y - X\hat{\Phi}) \ , \quad \bar{\nu} = T - k - n - 1 \ .^{11}$$

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 7.2).

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

[11] 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 $\bar{\nu} = T - k$.

Prior Specification Sometimes, priors are used as pure regularization devices. For example, we can shrink Φ to zero using a Lasso- or Ridge-prior (see Section 5.1). In the case of Ridge, we might specify $\Phi|\Sigma, \lambda \sim MN(\underline{\mu}, \underline{P}^{-1}, \Sigma)$ with $\underline{\mu} = 0$ and $\underline{P} = \lambda I$. Combining this with an improper prior for Σ, we get the MNIW posterior with

$$\bar{P} = \lambda I + X'X \ , \quad \bar{\mu} = \bar{P}^{-1}X'Y \ , \quad \bar{S} = (Y - X\hat{\Phi})'(Y - X\hat{\Phi}) \ , \quad \bar{\nu} = T - k - n - 1 \ .$$

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^* and $\hat{\Phi}^* = (X^{*'}X^*)^{-1}X^{*'}Y^*$. This is the posterior that would be obtained under an improper prior if our dummy observations (Y^*, X^*) were the true data. The resulting prior reflects the belief that (Φ, Σ) should be such that they generate observations Y^* and X^*.

Empirical Bayes The Ridge-prior above boils the hyperparameters $\underline{\mu}, \underline{P}, \underline{S}$ and $\underline{\nu}$ down to a single scalar, λ. Similarly, the Minnesota prior results in a 5-dimensional vector λ. Different values for λ lead to different models. As discussed in Sections 5.1 and 5.3, we could select the model (value of λ) that maximizes the MDD to trade off in-sample fit and model complexity obtain a good out-of-sample fit (forecasting performance). This is feasible in the case of the Ridge-prior, as the analytical expression for the MDD as a function of λ can be derived. It is more involved for the Minnesota prior, where the MDD is a non-trivial function of λ that we can only maximize numerically.

A more elegant solution is to use hierarchical Bayes modeling, as in Section 5.3. Combining

the Ridge-prior above with an improper prior for λ, we can derive

$$\begin{aligned} p(\lambda|Y,\Phi,\Sigma) &= p(\lambda|\Phi,\Sigma) \\ &\propto p(\Phi|\Sigma,\lambda)p(\lambda) \\ &\propto p(\Phi|\Sigma,\lambda) \\ &= (2\pi)^{-nk/2}|\Sigma|^{-k/2}|\lambda^{-1}I|^{-n/2}exp\left\{-\frac{1}{2}tr\left[\lambda\Sigma^{-1}\Phi'\Phi\right]\right\} \text{ ,}^{12} \end{aligned}$$

which let's us conclude that $\lambda|Y,\Phi,\Sigma \sim G\left(\frac{nk}{2}+1, \frac{1}{2}tr\left[\Sigma^{-1}\Phi'\Phi\right]\right)$. As a result, we can obtain the joint posterior $p(\Phi,\Sigma,\lambda|Y)$ using Gibbs sampling (see Section 8.2), as laid out in Algorithm 15.

Algorithm 15 (Gibbs Sampling: Hierarchical Ridge-VAR)**.**

1. *Initialize* λ^0 *(e.g. take* $\lambda^0 = 1$*).*
2. *For* $m = 1, 2, ..., M^*$*, given* λ^{m-1}*,*
   (a) *draw* (Φ^m, Σ^m) *from* $p(\Phi,\Sigma|Y,\lambda^{m-1})$*, i.e.*
      - *draw* Σ^m *from* $p(\Sigma|Y,\lambda^{m-1})$*,*
      - *draw* Φ^m *from* $p(\Phi|Y,\lambda^{m-1},\Sigma^m)$*,*

   (b) *draw* λ^m *from* $p(\lambda|Y,\Phi^m,\Sigma^m)$*.*

Under the Minnesota prior, we cannot derive the conditional posterior $p(\lambda|Y,\Phi,\Sigma)$ analytically. As a result, to obtain the joint the posterior $p(\Phi,\Sigma,\lambda|Y)$, we must use a numerical sampling algorithm like the RWMH- or SMC-algorithms (see Section 8.2). For this, we only need to be able to evaluate – up to proportionality – the prior $p(\Phi,\Sigma,\lambda) \propto p(\Phi|\Sigma,\lambda)$ and the likelihood.

10.2.2 Estimation of Structural VARs

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

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

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

[12] The first line follows by applying Bayes formula and using the fact that $p(Y|\Phi,\Sigma,\lambda) = (Y|\Phi,\Sigma)$. The second line is obtained because $p(\Sigma)$ is not a function of λ. The third line is obtained using $p(\lambda) \propto c$.

and HDs.

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 always write the set-/partially identified Φ_ε as

$$\Phi_\varepsilon = \Omega \Sigma_{tr} ,$$

whereby Σ_{tr} is the lower-triangular Cholesky-factor of Σ – i.e. it is s.t. $\Sigma = \Sigma_{tr}\Sigma_{tr}'$ –, and Ω is any orthogonal matrix – i.e. it is s.t. $\Omega\Omega' = I$. 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 Ω is not identified. From a Bayesian point of view, it means that any prior on Ω, $p(\Omega|\Phi, \Sigma, \cdot)$,[13]

[13] 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.

is not updated by the data:

$$\begin{aligned} p(\Phi, \Sigma, \Omega|Y) &\propto p(Y|\Phi, \Sigma, \Omega)p(\Phi, \Sigma, \Omega) \\ &= p(Y|\Phi, \Sigma)p(\Phi, \Sigma, \Omega) \\ &= p(Y|\Phi, \Sigma)p(\Phi, \Sigma)p(\Omega|\Phi, \Sigma, \cdot) \\ &\propto p(\Phi, \Sigma|Y)p(\Omega|\Phi, \Sigma, \cdot) \ . \end{aligned} \tag{10.8}$$

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. Any conclusions we draw for Ω come from our (conditional) prior for it!

As usual under Bayesian inference, a prior like $p(\Omega|\Phi, \Sigma, \cdot)$ 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.

Given this prior, we can obtain a draw from the posterior $p(\Phi, \Sigma, \Omega|Y)$ by first drawing (Φ, Σ) from their marginal posterior $p(\Phi, \Sigma|Y)$, derived in Section 10.2.1, and then, conditional on them, drawing Ω from the prior $p(\Omega|\Phi, \Sigma, \cdot)$. This is summarized in Algorithm 16. Of course, any draw $(\Phi^{(m)}, \Sigma^{(m)}, \Omega^{(m)})$ can be transormed into a draw $(\Phi^{(m)}, \Phi_\varepsilon^{(m)})$ by computing $\Phi_\varepsilon^{(m)} = \Sigma_{tr}^{(m)}\Omega^{(m)}$.

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

1. *Specify the prior $p(\Phi, \Sigma)$ and compute the marginal posterior $p(\Phi, \Sigma|Y)$ (inference for reduced-form VARs, as in Section 10.2.1).*
2. *Specify the prior $p(\Omega|\Phi, \Sigma, \cdot)$, and, 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.

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 (the ML estimator) is not unique (there is a set of values that all maximize the likelihood). Under an SVAR, any orthogonal Ω maximizes the likelihood, and – therefore – many Φ_ε maximize the likelihood. We can obtain the identified set of Φ_ε 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 on Φ_ε (on Ω); we draw Ω from a Uniform prior, while ensuring that the imposed restrictions are satisfied. As a result, frequentist inference for SVARs, as outlined in Algorithm 17, has a bit of a Bayesian flavor.

Algorithm 17 (Frequentist Inference in SVARs)**.**

1. *Compute* $(\hat{\Phi}, \hat{\Sigma})$ *(inference for reduced-form VARs, as in Section 10.2.1).*
2. *Specify* $p(\Omega|\Phi, \Sigma, \cdot)$ *– a Uniform distribution on the space of all orthogonal matrices restricted by identifying assumptions –, and, for* $m = 1 : M$

$$\textit{draw } \Omega^{(m)} \textit{ from } p(\Omega|\hat{\Phi}, \hat{\Sigma}, \cdot) \textit{ and compute } \Phi_\varepsilon^{(m)} = \hat{\Sigma}_{tr}\Omega^{(m)} .$$

The set of values $\{\Phi_\varepsilon^{(m)}\}_{m=1}^M$ *approximates the identified set of* Φ_ε *numerically.*

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 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 $(\hat{\Phi}, \hat{\Sigma})$. 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.

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, say, point- and sign-restrictions.[14] Given these restrictions that make up the prior $p(\Omega|\Phi, \Sigma, \cdot)$, one can use the algorithm proposed by Arias et al. (2018) to efficiently draw from $p(\Omega|\Phi, \Sigma, \cdot)$.[15]

[14]The 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. (10.4).

[15]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, point restrictions must be reflected in the prior directly because under a continuous distribution, the probability of obtaining a draw that exactly satisfies a point restriction is zero.

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

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

(recall Eq. (10.5)) 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

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

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.

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.

Further There are more approaches to reduce the identified set of Φ_ε. One method makes use of heteroskedasticity of shocks. Another method exploits non-Gaussianity of shocks.

10.3 State Space Models

Most if not all time series models can be written in so-called state space (SS) form as a system of two equations:

$$\begin{aligned} s_t &= \Phi(s_{t-1}, u_t; \theta) \ , \quad u_t \sim F_u(\theta) \ , \\ y_t &= \Psi(s_t; \theta) + \eta_t \ , \quad \eta_t \sim F_\eta(\theta) \ . \end{aligned} \tag{10.9}$$

The first equation is the transition equation. It deterimnes how a vector of generally unobserved state variables s_t evolves over time. The value of s_t today is a function of its value yesterday and an innovation u_t drawn from some distribution F_u.[16] The second equation

[16]The single lag-dynamics are w.l.o.g., as illustrated with the examples in Section 10.3.1.1. The intuition is provided by the companion form-VAR, which allows us to write a VAR with p lags as a VAR with a single

is the measurement equation. It determines how the vector of observed data at time t, y_t, relates to the vector of states at time t, s_t. Sometimes a vector of measurement errors η_t is added. It is typically assumed to enter additively and to be independent of innovations u_t. The goal is to estimate the parameters θ that determine the functions Φ and Ψ and distributions F_u and F_η. The functions Φ and Ψ can be nonlinear. In fact, they might not even be available in closed form, but we are only able to compute s_t numerically given s_{t-1}, u_t and θ.

We speak of a linear SS model if the functions Φ and Ψ are linear:

$$\begin{aligned} s_t &= \Phi_0 + \Phi_1 s_{t-1} + u_t \ , \\ y_t &= \Psi_0 + \Psi_1 s_t + \eta_t \ , \end{aligned} \tag{10.10}$$

where the innovations u_t and measurement errors η_t are WN processes. When they are Normally distributed, we say the SS model is linear and Gaussian. In this case, the parameter vector θ simply collects all the objects $\Phi_0, \Phi_1, \Psi_0, \Psi_1$ as well as the variances of u_t and η_t. Sometimes, these objects might be varying over time. For the characterization of a SS model as linear, this is not concerning, provided that we can write down for every period t a linear system of equations like above. In contrast, we get a nonlinear SS model if the function Φ is not linear in the states s_{t-1} and innovations u_t and/or if the function Ψ is not linear in the states s_t and the measurement errors η_t.

This section discusses the estimation of SS models, or, more precisely, the estimation of the parameters θ that determine the properties of a given SS model. We start with linear SS models in Section 10.3.1, going through examples, properties, so-called filtering- and smoothing-algorithms, as well as estimation. We then discuss analogous topics for nonlinear SS models in Section 10.3.2.

10.3.1 Linear State Space Models

10.3.1.1 Examples

(V)ARMA(p,q) The simplest example of the linear SS model from Eq. (10.10) is an AR(1), which we can write as

$$s_t = \phi s_{t-1} + u_t \ , \qquad y_t = s_t \ .$$

lag (see Section 10.2).

The MA(1) can be written as

$$\begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} u_t \ , \qquad y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} = s_{1t} \ .$$

The MA(q) can be written as

$$\begin{bmatrix} s_{1t} \\ \vdots \\ s_{q+1,t} \end{bmatrix} = \begin{bmatrix} 0 & I_q \\ 0 & 0' \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ \vdots \\ s_{q+1,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_q \end{bmatrix} u_t \ , \qquad y_t = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} s_{1t} \\ \vdots \\ s_{q+1,t} \end{bmatrix} = s_{1t} \ .$$

The AR(p) can be written as

$$\begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{pt} \end{bmatrix} = \begin{bmatrix} \phi_1 & \\ \phi_2 & I_{p-1} \\ \vdots & \\ \phi_p & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \\ \vdots \\ s_{p,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u_t \ , \qquad y_t = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{pt} \end{bmatrix} = s_{1t} \ .$$

Putting the two pieces together, we can write the ARMA(p, q) as

$$\begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{mt} \end{bmatrix} = \begin{bmatrix} \phi_1 & \\ \phi_2 & I_{m-1} \\ \vdots & \\ \phi_m & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \\ \vdots \\ s_{m,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{bmatrix} u_t \ , \qquad y_t = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{mt} \end{bmatrix} = s_{1t} \ ,$$

where $m = \max\{p, q+1\}$ and we write the ARMA(p, q) as an ARMA($m, m-1$) by setting $\phi_l = 0$ for $l = p+1 : m$ or $\theta_l = 0$ for $l = q+1 : m$.[17] Analogously, we can write the VARMA(p, q) as

$$\begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{mt} \end{bmatrix} = \begin{bmatrix} \Phi_1 & \\ \Phi_2 & I_{n(m-1)} \\ \vdots & \\ \Phi_m & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \\ \vdots \\ s_{m,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \Theta_1 \\ \vdots \\ \Theta_{m-1} \end{bmatrix} u_t \ , \qquad y_t = \begin{bmatrix} I & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{mt} \end{bmatrix} = s_{1t} \ .$$

[17] Note that the SS representation is not unique. For example, we can write the ARMA(1,1) also as

$$\begin{bmatrix} s_{1t} \\ s_{2t} \end{bmatrix} = \begin{bmatrix} \phi & \theta \\ 0 & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_t \ , \qquad y_t = s_{1t} \ .$$

It nests the VAR(p) and VMA(q) as special cases.

Linear Regression with Autocorrelated Errors Consider a linear regression model with ARMA(p, q)-disturbances:

$$y_t = x_t'\beta + v_t \quad \text{with} \quad v_t \sim \text{ARMA}(p, q) \ .$$

To write it in SS form, we use the above SS representation of an ARMA(p, q) and add to the measurement equation the term $x_t'\beta$. For example, under (p, q)=$(3, 1)$:

$$\begin{bmatrix} s_{1t} \\ s_{2t} \\ s_{3t} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 & 0 \\ \phi_2 & 0 & 1 \\ \phi_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} s_{1,t-1} \\ s_{2,t-1} \\ s_{3,t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta_1 \\ 0 \end{bmatrix} u_t \ , \qquad y_t = x_t'\beta + s_{1t} \ .$$

As a special case we have the linear regression without uncorrelated errors (i.e. with ARMA$(0, 0)$ errors), which we can also write in SS form:

$$s_t = u_t \quad , \ y_t = x_t'\beta + s_t \ .$$

Note that x_t is not a state variable because it is known to the researcher. It is not a measurement variable either because it is not an outcome variable of interest. The fact that it leads to a time-varying intercept in the measurement equation is not of further concern as this time-variation is known by the researcher and, therefore, in every period, the intercept can be treated as a constant.

Linear Regression with Time-Varying Parameters The SS model can also accommodate linear regressions with time-varying parameters (TVPs). For example, to allow for instability in the relationship between x_t and y_t, we might assume that parameters fluctuate around some mean values based on VAR(1)-dynamics:

$$s_t = \Phi_0 + \Phi_1 s_{t-1} + u_t \ , \quad y_t = x_t's_t + v_t \ .$$

This nests the case of independent AR(1) dynamics for each coefficient i in s_t: $s_{i,t} = \phi_0 + \phi_1 s_{i,t-1} + u_{it}$. Instead, to allow for a long-term trend in the relationship between x_t and y_t, we assume non-mean-reverting parameters, which can be modelled as random walks:

$$s_t = Is_{t-1} + u_t \ , \qquad y_t = x_t's_t + v_t \ .$$

In a similar vein, we can let the parameters in a VAR(p) vary over time. See for example Primiceri (2005); Del Negro and Primiceri (2015).

Under discrete steps for the TVPs, we speak of Markov-switching models. For example, with a single regressor in x_t, we could write

$$s_t \in \{\underline{s}, \bar{s}\} \sim \text{Markov chain w. transition matrix } Q = \begin{bmatrix} \underline{q} & 1-\underline{q} \\ 1-\bar{q} & \bar{q} \end{bmatrix} ,$$
$$y_t = x_t s_t + v_t .$$

This means that s_t can take on two values, $\underline{s}$ and $\bar{s}$, and its transition probabilities are $\mathbb{P}[s_t = \underline{s}|s_{t-1} = \underline{s}] = \underline{q}$ and $\mathbb{P}[s_t = \bar{s}|s_{t-1} = \bar{s}] = \bar{q}$. As a result, in some periods we have $y_t = x_t\bar{s} + v_t$, in others $y_t = x_t\underline{s} + v_t$. Because the relationship between x_t and y_t switches between two regimes, such models are also called Regime-Switching Models.

Factor Models A factor model relates the dynamics of an n-dimensional observable y_t to the dynamics of an r-dimensional vector of unobserved factors f_t, whereby $r << n$, i.e. the vector of factors is taken to be much lower dimensional than the vector of observables:

$$y_t = \Lambda f_t + \eta_t ,$$

where the $n \times r$ matrix Λ shows the loadings of the different series y_{it}, $i = 1 : n$, on the different factors f_{kt}, $k = 1 : r$. If f_t evolves according to, say, a VAR(p), we can write the factor model in SS form as

$$s_t = Fs_{t-1} + v_t , \qquad y_t = [\Lambda, 0, ..., 0]s_t + \eta_t , \tag{10.11}$$

where the $rp \times 1$ vectors $s_t = (f_t', ..., f_{t-p+1}')'$ and $v_t = (u_t', 0, ..., 0)'$ and the $rp \times rp$ matrix

$$F = \begin{bmatrix} \Phi_1 & ... & \Phi_{p-1} & \Phi_p \\ I & ... & 0 & 0 \\ \vdots & \ddots & & \vdots \\ 0 & & I & 0 \end{bmatrix}$$

implement the companion-form for the VAR(p) followed by f_t.

Mixed-Frequency Models Time series models that mix variables measured at different frequencies can also be written in SS form. For example, suppose we want to model the joint dynamics of interest rates and money supply, whereby interest rates are available at monthly frequency, whereas money supply is measured only every quarter. One approach is to aggregate the monthly interest rate data to quarterly frequency (i.e. to compute the average monthly interest rates observed over a quarter) and to model both series at quarterly frequency. Another approach is to write a SS model for interest rates and money supply at monthly frequency, whereby we observe the latter in every period (month), but the former only every three periods:

$$s_t = \Phi_0 + \Phi_1 s_{t-1} + u_t \ , \qquad \begin{matrix} y_{1t} = s_{1t} \ \text{ for } \ t = 1, 2, ..., T \\ y_{2t} = s_{2t} \ \text{ for } \ t = 3, 6, ..., \underline{T} \end{matrix} \ ,$$

where y_{1t} is the interest rate, y_{2t} is the stock of money supply in month t.[18] This approach might be preferred for several reasons.[19] See e.g. Schorfheide and Song (2015) for an example of a mixed frequency model.[20]

10.3.1.2 Properties

This section shows how to compute properties of s_t and y_t under generic SS models, including their mean, autocovariances, but possibly also IRFs, VDs and HDs. It is presumed that states s_t and observeables y_t are continuous random variables, but the analysis is readily extended to accommodate discrete s_t and/or y_t, as they appear in Markov-switching models, for example.

[18] Also, it is assumed that the third observation ($t = 3$) falls on the end of a quarter and therefore the observations for money supply go from $t = 3$ to the highest integer below T which is divisible by 3, denoted here by $\underline{T}$.

[19] The first is that forecasts can be updated not only every quarter – when a new observation for both the interest rate and money supply series becomes available – but every month, even within quarters, as soon as a new observation for the interest rate becomes available.

Second, by extracting the hidden states using filtering (see **??** below), we obtain an estimate for the money supply at monthly frequency, i.e. at higher frequency than available in the data.

Third, and related, forecasting such a variable that is measured officially at a higher frequency than it is modeled by the researcher is referred to as "nowcasting"; we forecast a variable that conceptually exists and has already realized, but an official datapoint will be released by the statistical agency only later and measured at a lower frequency.

[20] Note that money supply is a stock variable, meaning that we observe a snapshot of the monthly money supply series every three months (i.e. every quarter). For infrequently observed flow variables, the SS representation is slightly different because we observe every three months a sum or average of the past three monthly flows. For example, quarterly GDP is the sum of goods and services produced in the past three months, i.e. the sum of monthly GDP.

Recall the linear SS model from Eq. (10.10):

$$s_t = \Phi_0 + \Phi_1 s_{t-1} + u_t \ ,$$
$$y_t = \Psi_0 + \Psi_1 s_t + \eta_t \ .$$

Let $u_t \sim WN(0, \Sigma)$ and $\eta_t \sim WN(0, Q)$, with u_t and η_t uncorrelated, as is typically assumed.

In the case of a linear SS model, s_t just follows a VAR(1). Based on the analysis in Section 10.2, we know

$$\mathbb{E}[s_t] = (I - \Phi_1)^{-1}\Phi_0 \ , \quad \text{and} \quad \Gamma_h^s = \sum_{k=0}^{\infty} \Phi_1^{k+h} \Sigma \Phi_1^{k\prime} \ , \quad h \geq 0 \ .^{21}$$

In turn, using the measurement equation, we obtain

$$\mathbb{E}[y_t] = \Psi_0 + \Psi_1 \mathbb{E}[s_t] = \Psi_0 + \Psi_1 (I - \Phi_1)^{-1}\Phi_0 \ ,$$

and

$$\Gamma_h^y = \Psi_1 \Gamma_h^s \Psi_1' + \ \mathbf{1}\{h = 0\} Q \ .$$

Note that including one of the two intercepts Ψ_0 and Φ_0 suffices; w.l.o.g. one of the two can be set to zero, and in fact it would not possible to identify both from the data.[22] Also, based on the VMA(∞)- or GLP-representation of s_t, we can find that of y_t:

$$\begin{aligned} y_t &= \Psi_0 + \Psi_1 \left[\sum_{l=0}^{\infty} \Phi_1^l \Phi_0 + \sum_{l=0}^{\infty} \Phi_1^l u_{t-l} \right] + \eta_t \\ &= \Psi_0 + \Psi_1 (I - \Phi_1)^{-1}\Phi_0 + \eta_t + \Psi_1 \sum_{l=0}^{\infty} \Phi_1^l u_{t-l} \ . \end{aligned}$$

Note that the (main) source of dynamics in this model are innovations u_t. Measurement errors η_t just add some variability in y_t that is not tue do s_t and that is uncorrelated over time. Provided that the innovations u_t can be decomposed into shocks ε_t, one can undertake a causal analysis in the context of the more general (linear) SS model just as done for the

[21] Exact expressions for Γ_h^s are given by

$$vec(\Gamma_0^s) = [I - (\Phi_1 \otimes \Phi_1)]^{-1} vec(\Sigma) \ , \quad \text{and} \quad \Gamma_h^s = \Phi_1 \Gamma_{h-1}^s \ , \quad h \geq 1 \ .$$

[22] Of course, one can do that for each series y_{it} in y_t separately, as for some series it might be more convenient to specify $\Phi_{0,i}$, for others $\Psi_{0,i}$.

VAR in Section 10.2. For example, under $u_t = \Phi_\varepsilon \varepsilon_t$, one obtains the IRF

$$\frac{\partial y_{i,t+h}}{\partial \varepsilon_{j,t}} = \left[\Psi_1 \Phi_1^h \Phi_\varepsilon\right]_{ij} \ . \tag{10.12}$$

Similarly, one can compute also VDs and HDs. Note that for a VAR this expression boils down to Eq. (10.5).

10.3.1.3 Filtering & Smoothing

We write models in SS-form because some parts of s_t are unobserved. To conduct inference on θ, we need to account for this fact. Standard methods are (typically) not applicable; we cannot analytically derive the likelihood $p(Y|\theta)$. Furthermore, we may want to conduct inference on the unobserved s_t as a goal in itself.

Conditional on a value for the parameters θ, filtering algorithms compute the sequence of conditional distributions $p(s_t|Y_{1:t-1}, \theta)$, $p(y_t|Y_{1:t-1}, \theta)$ and $p(s_t|Y_{1:t}, \theta)$ for $t = 1 : T$. This allows us to form one-step ahead- and same-period-forecasts of s_t. It also allows us to compute the likelihood as $p(Y_{1:T}|\theta) = \prod_{t=1}^T p(y_t|y_{1:t-1}, \theta)$.[23] Algorithm 18 sketches the general idea behind filtering.

[23] However, note that, we can compute the likelihood for some SS models directly, without resorting to filtering. Examples from the class of linear SS models are AR and VAR processes, while an example from the class of nonlinear SS models are ARCH and GARCH processes. See Section 10.2 for the former and see Appendix for the latter.

Algorithm 18 (Generic Filtering Algorithm)**.**

1. *Initialize* $p(s_0|\theta)$.
2. *For* $t = 1 : T$, *given* $p(s_{t-1}|Y_{1:t-1}, \theta)$, [24]
   (a) *Forecast* s_t*: get*

$$p(s_t|Y_{1:t-1}, \theta) = \int p(s_t|s_{t-1}, Y_{1:t-1}, \theta)p(s_{t-1}|Y_{1:t-1}, \theta)ds_{t-1} \ .$$

   (b) *Forecast* y_t*: get*

$$p(y_t|Y_{1:t-1}, \theta) = \int p(y_t|s_t, Y_{1:t-1}, \theta)p(s_t|Y_{1:t-1}, \theta)ds_t \ .$$

   (c) *Update the forecast for* s_t *given the observation* y_t*:*

$$p(s_t|Y_{1:t}, \theta) = p(s_t|y_t, Y_{1:t-1}, \theta) = \frac{p(y_t|s_t, Y_{1:t-1}, \theta)p(s_t|Y_{1:t-1}, \theta)}{p(y_t|Y_{1:t-1}, \theta)} \ .$$

In case of a linear SS, we can (typically) characterize these integrals analytically, as shown below for a linear-Gaussian SS model.[25] For nonlinear SS models, we have to resort to numerical integration techniques, which leads to particle filtering-algorithms discussed below.

Filtering algorithms are "just-in-time-"algorithms, as they compute the forecasts of each s_t from the perspective of time $t-1$ or time t. Smoothing algorithms give us the densities $\{p(s_t|Y_{1:T}, \theta)\}_{t=1}^T$, i.e. the distribution of any s_t given the whole sample of data available, $Y_{1:T} = \{y_t\}_{t=1}^T$, i.e. given past, current (period t-) as well as future information. This allows us to form ex-post optimal forecasts of s_t. It also enables us to use a whole new class of estimation algorithms based on the idea of "data augmentation", whereby we treat the unobserved $\{s_t\}_{t=1}^T$ as parameters along with θ. (Note that $p(s_t|Y_{1:T}, \theta)$ is then the conditional posterior of a single s_t given θ.)

[24] For $t = 1$, this is simply $p(s_0|\theta)$.

[25] The discussion is instructive also for linear models with errors that follow other distributions than the Normal (e.g. the Markov-switching model presented in **??**) as analogous derivations can be made to get analytical expressions for the filtering- and subsequent smoothing-algorithms.

Filtering for Linear-Gaussian SS Models Under a linear-Gaussian SS model,

$$s_t = \Phi_0 + \Phi_1 s_{t-1} + u_t \ ,$$
$$y_t = \Psi_0 + \Psi_1 s_t + \eta_t \ ,$$

all of the densities in the filtering algorithm are Normal distributions. As such, they are characterized by their means and variances. Hence, computing the densities in Algorithm 18 above boils down to computing their means and variances.[26] Let $u_t \sim N(0, \Sigma)$ and $\eta_t \sim N(0, Q)$, with u_t and η_t uncorrelated, as is typically assumed.[27] To simplify notation, we drop $\theta = \{\Phi_0, \Phi_1, \Psi_0, \Psi_1, \Sigma, Q\}$ from the conditioning sets.

Algorithm 19 (Kalman Filter (Kalman, 1960))**.**

1. *Initialize* $p(s_0|\theta) = N(s_{0|0}, P_{0|0})$ *by specifying* $s_{0|0}$ *and* $P_{0|0}$.
2. *For* $t = 1 : T$, *given* $p(s_{t-1}|Y_{1:t-1}) = N(s_{t-1|t-1}, P_{t-1|t-1})$,
   (a) *Forecast* s_t*:*

$$p(s_t|Y_{1:t-1}) = N(s_{t|t-1}, P_{t|t-1}) \ , \quad \textit{with} \quad \begin{matrix} s_{t|t-1} = \Phi_0 + \Phi_1 s_{t-1|t-1} \\ P_{t|t-1} = \Phi_1 P_{t-1|t-1} \Phi_1' + \Sigma \end{matrix} \ .$$

   (b) *Forecast* y_t*:*

$$p(y_t|Y_{1:t-1}) = N(y_{t|t-1}, F_{t|t-1}) \ , \quad \textit{with} \quad \begin{matrix} y_{t|t-1} = \Psi_0 + \Psi_1 s_{t|t-1} \\ F_{t|t-1} = \Psi_1 P_{t|t-1} \Psi_1' + Q \end{matrix} \ .$$

   (c) *Update the forecast for* s_t *given the observation* y_t*:*

$$p(s_t|Y_{1:t}) = N(s_{t|t}, P_{t|t}) \ , \quad \textit{with} \quad \begin{matrix} s_{t|t} = s_{t|t-1} + P_{t|t-1} \Psi_1' F_{t|t-1}^{-1} (y_t - y_{t|t-1}) \\ P_{t|t} = P_{t|t-1} - P_{t|t-1} \Psi_1' F_{t|t-1}^{-1} \Psi_1 P_{t|t-1} \end{matrix} \ .^{28}$$

It is common to initialize the Kalman filter using the unconditional distribution of s_t: $s_{0|0} =$

[26] Similarly, other distributional families are also characterized by a set of parameters or moments. Thus, for many other linear SS models filtering boils down to computing the evolution of a set of parameters or moments over time, which can often be done analytically.

[27] The Kalman filter can easily be adjusted for the case when the two are correlated.

[28] This follows from the formula on turning a joint Normal into a conditional. Here we turn

$$\begin{bmatrix} s_t \\ y_t \end{bmatrix} \Big| Y_{1:t-1} \sim N\left(\begin{bmatrix} s_{t|t-1} \\ y_{t|t-1} \end{bmatrix} \ , \ \begin{bmatrix} P_{t|t-1} & P_{t|t-1}\Psi_1' \\ \Psi_1 P_{t|t-1} & F_{t|t-1} \end{bmatrix} \right) \quad \text{into} \quad s_t|y_t, Y_{1:t-1} = s_t|Y_{1:t} \ .$$

$\mathbb{E}[s_t] = (I - \Phi_1)^{-1}\Phi_0$ and $P_{0|0} = \mathbb{V}[s_t] = \Gamma_0^s$. However, this is only defined if s_t is stationary. For non-stationary s_t with mean zero, like unit root processes, one can use the conditional distribution of $s_0|(s_{-\tau} = 0)$ for some τ, which assumes the process was in equilibrium a couple of periods before the initialization. This implies that s_0 must be somewhere around that unconditional mean of zero, with higher variance under higher τ. Finally, it is also possible to treat $s_{0|0}$ and $P_{0|0}$ as parameters to estimate.

Based on the output of the Kalman filter, we obtain the log-likelihood

$$\begin{aligned}\log\ p(Y_{1:T}|\theta) &= \sum_{t=1}^{T} \log\ p(y_t|Y_{1:t-1}, \theta) \\ &= -\frac{nT}{2}\log(2\pi) - \frac{1}{2}\left(\sum_{t=1}^{T}\log|F_{t|t-1}|\right) - \frac{1}{2}\sum_{t=1}^{T}(y_t - y_{t|t-1})' F_{t|t-1}^{-1}(y_t - y_{t|t-1}) \ .\end{aligned}$$

We also obtain a series of one-step ahead forecasts of s_t, $\{s_t|Y_{1:t-1}\}_{t=1}^T$, as well as a series of same-period forecasts for s_t, $\{s_t|Y_{1:t}\}_{t=1}^T$. In order to get a prediction for s_t given the whole sample of data available, $Y_{1:T} = \{y_t\}_{t=1}^T$, i.e. given past, current (period t-) as well as future information, we use the so-called Kalman smoother. It yields the densities $\{p(s_t|Y_{1:T})\}_{t=1}^T$, and it is closely connected to the Carter and Kohn (1994) simulation smoother (henceforth CK Simulation Smoother), an algorithm to draw $\{s_t|Y_{1:T}\}_{t=1}^T$. For ease of exposition, the latter is presented first.

Smoothing for Linear-Gaussian SS Models The Simulation Smoother draws the sequence of states conditioning on all the data $Y_{1:T}$, $\{s_t|Y_{1:T}\}_{t=1}^T$. The density of interest can be written as

$$p(S_{1:T}|Y_{1:T}) = p(s_T|Y_{1:T})\prod_{t=1}^{T-1} p(s_t|s_{t+1}, Y_{1:T}) = p(s_T|Y_{1:T})\prod_{t=1}^{T-1} p(s_t|s_{t+1}, Y_{1:t}) \ .$$

From the Kalman filter, we know $p(s_T|Y_{1:T})$. Also, we know $p(s_t|Y_{1:t})$ as well as $p(s_{t+1}|, Y_{1:t})$, based on which we can construct the joint distribution $p(s_t, s_{t+1}|Y_{1:t})$, and, in turn, calculate the conditional $p(s_t|s_{t+1}, Y_{1:t})$. Once all these densities are obtained, we can draw $S_{1:T}$ sequentially by drawing first s_T, then s_{T-1} given s_T, etc. See Algorithm 20.

Algorithm 20 (CK Simulation Smoother (Carter and Kohn, 1994))**.**

1. *Run the Kalman filter to get* $\{s_{t|t}, s_{t|t-1}, P_{t|t}, P_{t|t-1}\}_{t=1}^{T}$.
2. *Draw* s_T^m *from* $s_T|Y_{1:T} \sim N(s_{T|T}, P_{T|T})$.
3. *For* $t = T-1, ..., 1$, *given draw* s_{t+1}^m *from* $s_{t+1}|s_{t+2}, Y_{1:t+1} \sim N(s_{t+1|t+2}, P_{t+1|t+2})$, *draw* s_t^m *from*

$$s_t|s_{t+1}, Y_{1:t} \sim N(s_{t|t+1}, P_{t|t+1}) \ , \quad \textit{with} \quad \begin{array}{l} s_{t|t+1} = s_{t|t} + P_{t|t}\Phi_1' P_{t+1|t}^{-1}(s_{t+1}^m - s_{t+1|t}) \\ P_{t|t+1} = P_{t|t} - P_{t|t}\Phi_1' P_{t+1|t}^{-1}\Phi_1 P_{t|t} \end{array} .$$

Based on the densities $\{p(s_t|s_{t+1}, Y_{1:t}) = N(s_{t|t+1}, P_{t|t+1})\}_{t=1}^{T}$ from the CK Simulation Smoother, the Kalman smoother constructs the densities $\{p(s_t|Y_{1:T}) = N(s_{t|T}^*, P_{t|T}^*)\}_{t=1}^{T}$ by applying the LIE (see Appendix).

Algorithm 21 (Kalman Smoother)**.**

1. *Run the Kalman filter to get* $\{s_{t|t}, s_{t|t-1}, P_{t|t}, P_{t|t-1}\}_{t=1}^{T}$.
2. *We know* $s_T|Y_{1:T} \sim N(s_{T|T}, P_{T|T})$.
3. *For* $t = T-1, ..., 1$, *given* $s_{t+1}|Y_{1:T} \sim N(s_{t+1|T}^*, P_{t+1|T}^*)$, *we get*

$$s_t|Y_{1:T} \sim N(s_{t|T}^*, P_{t|T}^*) \ , \quad \textit{with} \quad \begin{array}{l} s_{t|T}^* = s_{t|t} + P_{t|t}\Phi_1' P_{t+1|t}^{-1}(s_{t+1|T}^* - s_{t+1|t}) \\ P_{t|T}^* = P_{t|t} + P_{t|t}\Phi_1' P_{t+1|t}^{-1}(P_{t+1|T}^* - P_{t+1|t})P_{t+1|t}^{-1}\Phi_1 P_{t|t} \end{array} .$$

Both smoothing algorithms need to be adjusted for cases in which the RV $s_t|s_{t+1}$ contains deterministic elements. This happens for example when s_t is a VAR(p) (or AR(p)) written in companion-form as a VAR(1) (see Section 10.2 and Section 9.2). Adjustments for this case are discussed in the Appendix.

10.3.1.4 Estimation

The linear SS model consists of two linear regressions:

$$\begin{aligned} s_t &= \Phi_0 + \Phi_1 s_{t-1} + u_t \ , \quad \mathbb{V}[u_t] = \Sigma \ , \\ y_t &= \Psi_0 + \Psi_1 s_t + \eta_t \ , \quad \mathbb{V}[\eta_t] = Q \ . \end{aligned}$$

Thereby, $Y_{1:T} = \{y_t\}_{t=1}^T$ is observed. Hence, the only obstacle to estimating the parameters of interest, $\theta = \{\Phi_0, \Phi_1, \Psi_0, \Psi_1, \Sigma, Q\}$, is the fact that the states $S_{1:T} = \{s_t\}_{t=1}^T$ are not observed. It is useful to distinguish between the "complete-data likelihood" $p(Y_{1:T}, S_{1:T}|\theta)$, which we could form if we knew $S_{1:T}$, and the "incomplete-data likelihood" $p(Y_{1:T}|\theta)$, which only depends on data $Y_{1:T}$.

Sometimes, knowing $S_{1:T}$ is not necessary to write down and evaluate the likelihood $p(Y_{1:T}|\theta)$ analytically. Examples are AR(p) and VAR(p) models as well as linear regressions with autocorrelated errors. In some of these cases, we might even obtain an analytical frequentist estimator $\hat{\theta}$ or analytically derive the posterior $p(\theta|Y_{1:T})$, either directly or by iterating on analytical conditional estimators (see Meng and Rubin (1993) algorithm in Section 8.1) or analytical conditional posteriors (see Gibbs sampling algorithm in Section 8.2). In many cases, however, no analytical expressions for point estimators or posteriors are available.

In this general case, we can evaluate $p(Y_{1:T}|\theta)$ using a filtering algorithm. Intuitively, filtering algorithms construct this incomplete-data likelihood by integrating out the hidden states from the complete-data likelihood. In the case of a linear SS model, evaluating the likelihood using filtering is typically computationally very efficient because the algorithm just iterates on a set of analytical expressions. In addition, linear SS models permit smoothing algorithms that allow us to estimate or draw the hidden states $\{s_t\}_{t=1}^T$ conditional on all data $Y_{1:T}$. Estimation approaches for linear SS models make use either of this extraction of hidden states (referred to as "data augmentation") or of the fact that we can evaluate the likelihood in a computationally efficient way.

Frequentist Inference We can get $\hat{\theta}$ by numerically maximizing the (log-)likelihood obtained by the Kalman filter. As for any numerical optimization problem, a downside of this approach is that – in general – we have no guarantee that the optimum we find is indeed the global- as opposed to a local optimum.

A more efficient and numerically stable alternative is oftentimes the Expectation-Maximization (EM) algorithm. It iteratively constructs and maximizes the expectation of the complete-data likelihood,

$$\mathbb{E}[p(Y_{1:T}, S_{1:T}|\theta)] = p(Y_{1:T}, \{s^*_{t|T}\}_{t=1}^T|\theta) \ ,$$

by replacing the unknown states $S_{1:T} = \{s_t\}_{t=1}^T$ by the estimates $\{s^*_{t|T}\}_{t=1}^T = \{\mathbb{E}[s_t|Y_{1:T}, \theta]\}_{t=1}^T$ obtained by the Kalman smoother.

Algorithm 22 (Expectation-Maximization (EM) Algorithm)**.**

1. *Initialize* θ^0.
2. *For* $m = 1, 2, ...,$
   (a) *(Expectation) given* θ^{m-1}*, run the smoothing algorithm to obtain* $\{s^*_{t|T}\}_{t=1}^T$.
   (b) *(Maximization) given* $\{s^*_{t|T}\}_{t=1}^T$*, find*

$$\theta^m = \arg\max_{\theta\in\Theta} \log\ p(Y_{1:T}, \{s^*_{t|T}\}_{t=1}^T|\theta)\ .$$

   (c) *if* $||(\theta^{m+1} - \theta^m|| < \varepsilon$*, take* $\hat{\theta} = \theta^{m+1}$*. Else, proceed to the next iteration.*

The EM algorithm works analogously to the Meng and Rubin (1993) procedure presented in Section 8.1: it iterates on two conditional estimators, $S^*_{1:T}|\theta, Y_{1:T}$ and $\hat{\theta}|S_{1:T}, Y_{1:T}$, to find the joint estimator $(\hat{\theta}, S^*_{1:T})|Y_{1:T}$. The hidden states $S_{1:T}$ are treated as parameters along with θ, whereby, of course, in the end we care mostly about the estimator of θ.

The expectation step is typically performed very fast, as running the smoothing algorithm just involves iterating on a set of analytical expressions. If, in addition, we can solve the optimization problem in the maximization-step analytically, this results in a very efficient procedure to find $\hat{\theta}$.[29] Provided that the maximization step can be done analytically, the EM algorithm is typically more efficient than numerically maximizing the likelihood. In contrast, as soon as the estimator of (a subset of) θ has to be found numerically, we might be better off maximizing the likelihood numerically with respect to all parameters in θ at once. Examples of models easily estimated using the EM algorithm include mixed-frequency VARs or linear regressions with regime-switching parameters (see Section 10.3.1.1).

[29]Typically, we can divide the complete-data likelihood into two components, each with its distinct set of parameters, which effectively means that we are estimating two models, one related to the measurement equation, one to the transition equation. Concretely,

$$p(Y_{1:T}, S_{1:T}|\theta) = \prod_{t=1}^T p(y_t, s_t|Y_{1:t-1}, S_{1:t-1}, \theta) = \prod_{t=1}^T p(y_t|s_t, \theta)p(s_t|s_{t-1}, \theta) = \left[\prod_{t=1}^T p(y_t|s_t, \theta)\right]\left[\prod_{t=1}^T p(s_t|s_{t-1}, \theta)\right]\ ,$$

whereby Φ_0, Φ_1 and Σ appear only in the second term, and Ψ_0, Ψ_1 and Q appear only in the first term.

If needed, we can break down the maximization step into further conditional maximization steps. This is useful if $\hat{\theta}|S_{1:T}, Y_{1:T}$ is not available analytically, but we can find analytical estimators for subsets of θ given the rest of θ.

Bayesian Inference To implement Bayesian inference, we also have two approaches available. They mimic the two frequentist estimation methods.

To obtain a particle approximation $\{\theta^i, W^i\}_{i=1}^N$ of the posterior $p(\theta|Y_{1:T})$, we can use numerical posterior sampling techniques like the Metropolis Hastings (MH) or Sequential Monte Carlo (SMC) algorithms (see Section 8.2). Their only requirement is that we are able to evaluate the likelihood and prior, which makes them very widely applicable. A downside is that these algorithms can be slow to converge.

Sometimes – in particular under slower likelihood-evaluations –, a more efficient alternative is the Carter and Kohn (1994)-Gibbs sampler. In line with the general Gibbs sampling-algorithm (see Section 8.2) and analogous to the frequentist EM algorithm, it iterates on the two conditional posteriors $p(S_{1:T}|Y_{1:T}, \theta)$ and $p(\theta|Y_{1:T}, S_{1:T})$ to find the joint posterior $p(\theta, S_{1:T}|Y_{1:T})$.

Algorithm 23 (Carter and Kohn (1994) Gibbs Sampler)**.**

1. *Initialize θ^0, e.g. by drawing from the prior $p(\theta)$.*
2. *For $m = 1, 2, ...,$*
   (a) given θ^{m-1}, run the smooting algorithm to obtain a draw $S_{1:T}^m \sim p(S_{1:T}|Y_{1:T}, \theta^{m-1})$.
   (b) given $S_{1:T}^m$, draw θ^m from the (complete-data) posterior

$$p(\theta|Y_{1:T}, S_{1:T}^m) \propto p(Y_{1:T}, S_{1:T}^m|\theta)p(\theta) \ ,$$

obtained using the (complete-data) likelihood $p(Y_{1:T}, S_{1:T}^m|\theta)$.

Obtaining a draw from $p(S_{1:T}|Y_{1:T}, \theta)$ tends to be fast as it involves (i) iterating on a set of analytical expressions to obtain conditional distributions and (ii) iteratively drawing from them (see e.g. the CK Simulation Smoother, obtained under a linear-Gaussian SS model). If we can also analytically derive the posterior $p(\theta|Y_{1:T}, S_{1:T}^m)$, this results in a very efficient procedure to obtain a set of draws from $p(\theta, S_{1:T}|Y_{1:T})$ and, therefore, $p(\theta|Y_{1:T})$. Thereby, we can again break θ down into several subsets for which conditional posteriors are available given the rest of θ.

10.3.2 Nonlinear State Space Models

10.3.2.1 Examples

Models of Volatility Dynamics An example of the nonlinear SS model from Eq. (10.9) is an AR(1) model with stochastic volatility (SV):

$$y_t = \phi y_{t-1} + u_t \ , \quad \text{with} \quad u_t \sim N(0, e^{h_t}) \ , \quad h_t = \gamma h_{t-1} + e_t \ , \quad e_t \sim N(0,1) \ .$$

We can write it in SS form as

$$\begin{matrix} s_{1,t} = \phi s_{1,t-1} + \sqrt{e^{s_{2t}}}\, v_t \\ s_{2,t} = \gamma s_{2,t-1} + e_t \end{matrix} \ , \quad (v_t, e_t)' \sim N(0, I) \ , \qquad y_t = s_{1t} \ . \tag{10.13}$$

In a similar vein, we can include SV into a VAR(p).

Rather than assuming that volatility varies based on an exogenous process, we might assume that it varies based on past values of the observed series y_t. Such conditional heteroskedasticity models are particularly popular in finance. At their most basic form, they postulate that an extreme realization of y_t (e.g. an extreme return of an asset) tends to be followed by further extreme realizations. An example is the Autoregressive Conditional Heteroskedasticity (ARCH) model of order 1:

$$y_t \sim N(0, h_t) \ , \quad h_t = \omega + \alpha y_{t-1}^2 \ ,$$

or the Generalized ARCH (GARCH) model of order (1,1):

$$y_t \sim N(0, h_t) \ , \quad h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1} \ . \tag{10.14}$$

In SS form, the ARCH(1) process yields

$$\begin{matrix} s_{1,t} = \sqrt{s_{2t}}\, u_t \\ s_{2,t} = \omega + \alpha s_{1,t-1}^2 \end{matrix} \ , \quad u_t \sim N(0,1) \ , \qquad y_t = s_{1t} \ ,$$

while for the GARCH(1,1) we get

$$\begin{matrix} s_{1,t} = \sqrt{s_{2t}}\, u_t \\ s_{2,t} = \omega + \alpha s_{1,t-1}^2 + \beta s_{2,t-1} \end{matrix} \ , \quad u_t \sim N(0,1) \ , \qquad y_t = s_{1t} \ .$$

Note that SV- as well as conditional heteroskedasticity-models are conditionally linear, i.e. the transition and measurement equations are linear if we condition on past states, s_{t-1}.

This can facilitate their estimation (see Section 10.3.2.4).

Time-Varying Parameter Models Some TVP models also lead to nonlinear SS representations. For example, consider the following regime-switching model:

$$\begin{matrix} s_{1t} = \phi_1 s_{1,t-1} \ u_{1t} \\ s_{2t} = \phi_2 s_{2,t-1} \ u_{2t} \end{matrix} \ , \qquad y_t = s_{1t} + \delta \ \mathbf{1}\left\{s_{2t} > 0\right\} \ .$$

Here, s_{1t} governs the AR(1) dynamics in y_t, while s_{2t} is the latent regime that determines whether the mean of y_t is zero or equal to δ in any period t. A nonlinear SS model is also obtained if the transition probabilities in Markov-switching models are time-varying. Building on the example above, say

$$s_t \in \left\{\underline{s}, \bar{s}\right\} \ , \qquad y_t = x_t s_t + v_t \ ,$$

with transition probabilities modeled using a probit model; e.g. $\mathbb{P}[s_t = \underline{s} | s_{t-1} = \underline{s}] = \Phi(\underline{c} + z_t' \underline{\beta})$, where $\Phi(\cdot)$ is the cdf of a standard Normal distribution, and similarly for $\mathbb{P}[s_t = \bar{s} | s_{t-1} = \bar{s}]$. By adding a second state variable, these probabilities can be duration-dependent:

$$s_{1t} \in \left\{\underline{s}, \bar{s}\right\} \sim \text{ Markov chain } , \quad s_{2t} = \begin{cases} s_{2t} + 1 & \text{if } s_{1t} = s_{1,t-1} \\ 1 & \text{else} \end{cases} \ ,$$

and e.g. $\mathbb{P}[s_{1t} = \underline{s} | s_{1,t-1} = \underline{s}, s_{2,t-1}] = \Phi(\underline{c} + z_t' \underline{\beta} + \gamma s_{2,t-1})$. Here, $s_{2,t-1}$ measures how long $s_{1,t-1}$ has been in the same state.

Nonlinear DSGE Models & Beyond There are endless further examples of models that yield nonlinear SS representations. The nonlinear SS model is even general enough to accommodate cases for which the functions generating the dynamics are not available in closed form, but one numerically obtains s_t given s_{t-1} and u_t as well as some parameter values θ that determine the exact mapping. For example, given values for parameters like risk aversion, discount rate or persistence of a productivity process and given a series of shocks u_t, a macroeconomic model can generate observations for variables like consumption, GDP, interest rates, etc., without us knowing explicitly the function that generates these data. An exception are macro models solved using perturbation methods. For example, a second-order linearization around steady state gives a quadratic function Φ, while a first-order linearization gives a linear function Φ, both of which can be derived in closed form. The Appendix contains a brief discussion of the estimation of the Real Business Cycle (RBC) model, the predecessor of all modern DSGE models.

10.3.2.2 Properties

Recall the nonlinear SS model from Eq. (10.9):

$$\begin{aligned} s_t &= \Phi(s_{t-1}, u_t; \theta) \ , \quad u_t \sim F_u(\theta) \ , \\ y_t &= \Psi(s_t; \theta) + \eta_t \ , \quad \eta_t \sim F_\eta(\theta) \ . \end{aligned}$$

The moments of s_t and y_t can be computed numerically, by simulaton. Given a simulated series $\{\tilde{s}_t, \tilde{y}_t\}_{t=1}^{T_s}$, we obtain

$$\mathbb{E}[y_t] \approx \frac{1}{T_s} \sum_{t=1}^{T} \tilde{y}_t \quad \text{and} \quad \Gamma_h^y \approx \frac{1}{T_s - h} \sum_{t+h}^{T_s} (\tilde{y}_t - \hat{\mu}_y) (\tilde{y}_{t-h} - \hat{\mu}_y)' \ ,$$

provided the series y_t is stationary. We can approximate $\mathbb{E}[s_t]$ and Γ_h^s analogously.

IRFs in nonlinear SS models are also computed numerically (provided that innovations u_t can be decomposed into shocks ε_t). In addition, in nonlinear models, the impulse-response of y_{t+h} to a shock ε_{jt} can depend on the state s_{t-1} (and therefore possibly on the time period we are in) as well as on the shocks that occur between period t and $t+h$. See for example Aruoba et al. (2022), where s_{t-1} affects whether the economy is at or away from the effective lower bound on interest rates. Hence, we take a step back and define the object of interest as

$$IRF_{t,ij}^h \equiv \mathbb{E}[y_{i,t+h} | s_{t-1}, \varepsilon_{jt} = 1] - \mathbb{E}[y_{i,t+h} | s_{t-1}] \ , \tag{10.15}$$

i.e. the difference between the expected value of y_{t+h} that would be obtained starting from s_{t-1} and under $\varepsilon_{jt} = 1$ and its expected value starting from s_{t-1} that does not condition on $\varepsilon_{jt} = 1$. The expectation operator signifies that, in the latter, we average out the distribution of all shocks in all periods, while in the former we average out the distribution of all shocks in all periods except for ε_{jt}, which is set to one. For a linear (SS) model, averaging out the shocks is equivalent to setting them to their mean values of zero. The above equation then simplifies to the expression $\left[\Psi_1 \Phi_1^h \Phi_\varepsilon\right]_{ij}$ from Eq. (10.12), which is not state-dependent.[30] For the nonlinear SS model, the IRF can be computed numerically by approximating the expectations in Eq. (10.15) as outlined in Algorithm 24.

[30] Concretely, we get $\mathbb{E}[y_{t+h} | s_{t-1}, \varepsilon_{jt} = 1] = \Psi_1 \Phi_1^{h+1} s_{t-1} + \Psi_1 \Phi_1^h \Phi_\varepsilon e_j$ and $\mathbb{E}[y_{t+h} | s_{t-1}] = \Psi_1 \Phi_1^{h+1} s_{t-1}$, where e_j is a vector of zeros with a one in position j.

Algorithm 24 (Numerical Computation of IRFs)**.**

1. *For* $m = 1 : M$*,*
   (a) *draw* $\{\tilde{\varepsilon}_t^m\}_{t=1}^h$ *from the distribution of* ε_t*.*[31]
   (b) *given* s_{t-1}*, use* $\{\tilde{\varepsilon}_t^m\}_{t=1}^h$ *to construct the simulated series* $\{\tilde{s}_t^m\}_{t=1}^h$ *and* $\{\tilde{y}_t^m\}_{t=1}^h$*.*
   (c) *take* $\{\tilde{\varepsilon}_t^m\}_{t=1}^h$*, increase the value of* ε_t^m *by* 1*, and write the resulting series of shocks as* $\{\check{\varepsilon}_t^m\}_{t=1}^h$*.*
   (d) *given* s_{t-1}*, use* $\{\check{\varepsilon}_t^m\}_{t=1}^h$ *to construct the simulated series* $\{\check{s}_t^m\}_{t=1}^h$ *and* $\{\check{y}_t^m\}_{t=1}^h$*.*
2. *The desired impulse-response* $IRF_{t,ij}^h$ *is approximated numerically by*

$$\widehat{IRF}_{t,ij}^h = \frac{1}{M}\sum_{m=1}^{M} \check{y}_{t+h}^m - \frac{1}{M}\sum_{m=1}^{M} \tilde{y}_{t+h}^m = \frac{1}{M}\sum_{m=1}^{M} \left(\check{y}_{t+h}^m - \tilde{y}_{t+h}^m\right) .$$

In principle, a VD can also be computed numerically: one can simulate the series y_t letting only shock j vary, compute its variance and compare it to the variance obtained when all shocks are set to zero, including shock j. However, in a nonlinear model, the contributions of all shocks do not necessarily add up to the overall variance of y_t. The same holds for the HD: the hypothetical evolutions of y_t if only one of the shocks is set to its actual (estimated) evolution and all others are set to zero do not necessarily add up to the overall evolution of y_t observed in the data.

Note that these approaches to numerically approximate the IRF, VD and HD can also be applied for linear models. This is an attractive alternative if the analytical expressions are cumbersome to deal with.

[31] Shocks are typically defined to be i.i.d. random variables and, w.l.o.g., their variance is set to 1. Usually, they are assumed to be Normal, yielding overall $\varepsilon_{jt} \overset{i.i.d.}{\sim} N(0,1)$. Based on shocks ε_t, one can compute innovations u_t and simulate the series s_t and y_t.

However, instead of writing the model in terms of innovations u_t and then decomposing them into shocks ε_t using certain identification assumptions, some nonlinear SS models, like those obtained from nonlinear DSGE models, directly specify the dynamics of s_t as a function of shocks ε_t.

10.3.2.3 Filtering

For nonlinear SS models, we can evaluate the likelihood and extract the hidden states using particle filtering approaches, which numerically approximate the densities and integrals in Algorithm 18. Recall the general nonlinear SS model,

$$\begin{aligned} s_t &= \Phi(s_{t-1}, u_t; \theta) \ , \quad u_t \sim F_u(\theta) \ , \\ y_t &= \Psi(s_t; \theta) + \eta_t \ , \quad \eta_t \sim F_\eta(\theta) \ . \end{aligned}$$

The simplest version of a particle filter is the bootstrap particle filter (BSPF). For a given θ, which determines the functions Φ and Ψ and distributions F_u and F_η, the BSPF proceeds as follows.

Algorithm 25 (Bootstrap Particle Fitler)**.**

1. *Initialization: draw particles* $s_0^m \sim p(s_0|\theta)$ *and set* $W_0^m = 1$ *for* $m = 1 : M$*. Trivially,* $\{s_0^m, W_0^m\}_{m=1}^M$ *approximate* $p(s_0|\theta)$*.*
2. *For* $t = 1 : T$*, given* $\{s_{t-1}^m, W_{t-1}^m\}_{m=1}^M$ *that approximate* $p(s_{t-1}|Y_{1:t-1})$*,*
   (a) *Forecast* s_t*: for* $m = 1 : M$*, draw* $u_t^m \sim F_u(\theta)$ *and compute* $\tilde{s}_t^m = \Phi(s_{t-1}^m, u_t^m; \theta)$*.*
   (b) *Forecast* y_t*: define the incremental weights* $\tilde{w}_t^m = p(y_t|\tilde{s}_t^m; \theta)$ *and approximate the incremental likelihood by*

$$\hat{p}(y_t|Y_{1:t-1}, \theta) = \frac{1}{M} \sum_{m=1}^{M} \tilde{w}_t^m W_{t-1}^m \ .$$

   (c) *Update: define normalized weights* $\tilde{W}_t^m = \dfrac{\tilde{w}_t^m W_{t-1}^m}{\hat{p}(y_t|Y_{1:t-1}, \theta)}$*.*
   (d) *Selection (optional): resample particles by drawing* $\{\tilde{s}_t^m\}_{m=1}^M$ *from* $\{\tilde{s}_{t-1}^m, \tilde{W}_{t-1}^m\}_{m=1}^M$ *and setting* $W_t^m = 1 \ \forall \ m$*. Else set* $s_t^m = \tilde{s}_t^m$ *and* $W_t^m = \tilde{W}_t^m \ \forall \ m$*.*[32]

Given estimates $\hat{p}(y_t|Y_{1:t-1}, \theta)$ for $t = 1 : T$, we approximate the log-likelihood as

$$\log \ \hat{p}(Y_{1:T}|\theta) = \sum_{t=1}^{T} \log \ \hat{p}(y_t|Y_{1:t-1}, \theta) \ .$$

The BSPF only works (well) if there are measurement errors η_t (with a high enough vari-

[32] Analogously as in the SMC algorithm in the Appendix to Chapter 8, the selection step can be executed if the effective sample size falls below a certain threshold: e.g. $ESS_t < M/2$, where $ESS_t = M/\left(\frac{1}{M}\sum_{m=1}^{M}(\tilde{W}_t^m)^2\right)$.

ance). Without measurement errors, $p(y_t|s_t,\theta)$ is a pointmass at $y_t = \Psi(s_t, u_t; \theta)$. Under a continuous distribution for u_t, the probability of obtaining exactly the "true" value $\tilde{s}_t^m = s_t$ that yields the observed y_t is zero, which means that after one iteration all particles would have $\tilde{w}_t^m = 0$ and the algorithm would break down.[33] More efficient particle filters incorporate information on where y_t lies when computing $\tilde{s}_t^m$ given $\tilde{s}_{t-1}^m$. This leads to the following algorithm for a generic particle filter.

Algorithm 26 (Generic Particle Filter)**.**

Replace the following two steps in Algorithm 25:

(a) Forecast s_t: draw $\tilde{s}_t^m$ from $g_t(\tilde{s}_t^m|s_{t-1}^m,\theta)$, and define the importance weights

$$w_t^m = \frac{p(\tilde{s}_t^m|s_{t-1}^m,\theta)}{g_t(\tilde{s}_t^m|s_{t-1}^m,\theta)} \ .$$

(b) Forecast y_t: define the incremental weights $\tilde{w}_t^m = p(y_t|\tilde{s}_t^m,\theta)w_t^m$ and approximate the incremental likelihood, as before, by

$$\hat{p}(y_t|Y_{1:t-1},\theta) = \frac{1}{M}\sum_{m=1}^{M} \tilde{w}_t^m W_{t-1}^m \ .$$

The density $g_t(\tilde{s}_t^m|s_{t-1}^m,\theta)$ is supposed to incorporate information from y_t and, therefore, to yield draws $\tilde{s}_t^m$ that are compatible with y_t. For more details on how to choose g_t and more details on particle filtering, see Herbst and Schorfheide (2015).

10.3.2.4 Estimation

For some nonlinear SS models, we can derive the likelihood $p(Y_{1:T}|\theta)$ analytically. This is the case, for example, for GARCH models (see Appendix). In such cases, we can numerically maximize the likelihood or approximate the posterior using posterior sampling methods. These approaches work well since the likelihood can be evaluated instantaneously.

In other cases, we can evaluate the likelihood using the particle filter. In principle, these likelihood evaluations could then be used to numerically maximize the likelihood or approximate the posterior. However, the stochastic likelihood-evaluation via particle filtering introduces noise in the estimation process and tends to be computationally rather costly.

[33] The Kalman filter does not have this problem because in the linear-Gaussian SS model, we can compute $p(y_t|y_{t-1};\theta)$, i.e. we can go from s_t to y_t over y_{t-1} rather than going directly from s_t to y_t.

For Bayesian inference, this is not a big concern; the particle approximation of the posterior tends to work well despite the noise in the likelihood evaluation. Nevertheless, computational considerations become important; the SMC algorithm is preferred to the MH algorithm due to the possibility to parallelize computations in the former. Nonlinear DSGE models are often estimated using the resulting, so-called SMC^2-approach that embeds likelihood evaluations via particle filtering into the SMC sampler (see Appendix and see Herbst and Schorfheide (2015)).

In contrast, for frequentist inference, the slow likelihood evaluations via particle filtering typically mean that numerical maximization of the likelihood is not an option. In these cases we can rely on Bayesian techniques and approximate the ML estimator as the posterior mode under a uniform or improper prior. When the posterior is obtained via a particle approximation $\{\theta^i, W^i\}_{i=1}^N$, we can aproximate the ML estimator as the Maximum A-Posteriori (MAP) estimator: the draw θ^i from the particle approximation that corresponds to the highest value of the posterior.

As in the case of linear SS models, a preferred alternative is typically to rely on data augmentation, i.e. use the EM algorithm or the Carter and Kohn (1994)-Gibbs sampler. Despite the fact that a smoothing algorithm is not available for nonlinear SS models, we can oftentimes break up the nonlinear SS model into several linear models by breaking up the unknown parameters and states $(\theta, S_{1:T})$ into sub-groups and carrying out the estimation by iterating on the different sub-parameters and -states, conditioning on all others. This frequently involves finding smart ways to rewrite equations. An instructive example is the AR(1) with SV from Eq. (10.13) as estimated by Kim et al. (1998):

$$\begin{aligned} s_{1,t} &= \phi s_{1,t-1} + \sqrt{e^{s_{2t}}}\, v_t \\ s_{2,t} &= \gamma s_{2,t-1} + e_t \end{aligned} \quad , \quad (v_t, e_t)' \sim N(0, I) \; , \qquad y_t = s_{1,t} \; .$$

Conditioning on the volatility process $\{s_{2,t}\}_{t=1}^T$, we have a linear SS model, which consists of i) an AR(1) model with known heteroskedasticity for $s_{1,t} = y_t$ and ii) an AR(1) model with homoskedastic errors for $s_{2,t}$. This allows us to estimate ϕ and γ. Conditional on (ϕ, γ), we can arrive at a linear SS model that allows us to estimate $\{s_{2,t}\}_{t=1}^T$ as follows. First, note that $\tilde{y}_t \equiv s_{1,t} - \phi s_{1,t-1} = y_t - \phi y_{t-1} \sim N(0, e^{s_{2,t}})$. In turn, we can write this as $\tilde{y}_t = e^{\frac{1}{2}s_{2,t}} u_t$ with $u_t \sim N(0,1)$. Squaring both sides and taking logs, we arrive at a linear measurement equation,

$$y_t^* = s_{2,t} + u_t^* \; ,$$

where $y_t^* = \log(\tilde{y}_t^2)$ is known and $u_t^* = \log(u_t^2)$ is the logarithm of a χ_1^2-distributed RV, u_t^2. Togher with the linear transition equation for $s_{2,t}$ from above, this defines a linear SS model,

whereby γ is known and the goal is to obtain an estimate for the hidden volatility process $\{s_{2,t}\}_{t=1}^{T}$. To be able to do so efficiently using the CK Simulation Smoother, Kim et al. (1998) approximate the distribution of u_t^* using a mixture of Normals.

Appendix

Estimation of VARs

Minnesota Prior A common view is that, in absence of shocks, macroeconomic variables are expected to stay at their current levels, i.e. they follow random walks. The Minnesota prior constructs dummy observations to center the distribution of Φ at values that imply a random walk for y_t. It is popular for several reasons. First, it is easily scalable when more lags or variables are added into the analysis. Second, it corrects for the fact that the MLE tends to be downward biased in small samples if y_t has roots near or at unity. Third, the hyperparameter λ is interpretable. When $\lambda = 0$, all observations are zero and we get an improper prior. The larger λ, the more informative the prior. As elaborated on in Section 5.1, λ is often selected to maximize the MDD and therefore the one-period ahead prediction ability of the model.

Let $\underline{y}$ and $\underline{s}$ be the $n \times 1$ vectors of means and standard deviations of y_t, and let λ be a 5×1 vector of hyperparameters. To illustrate, consider a bivariate VAR(3). Extensions to other cases are straightforward. The dummy observations Y^* and X^* are constructed in three chunks.

The first np observations are:

$$\begin{bmatrix} \lambda_1 \underline{s}_1 & 0 \\ 0 & \lambda_1 \underline{s}_2 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \lambda_1 \underline{s}_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 \underline{s}_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 \underline{s}_1 2^{\lambda_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 \underline{s}_2 2^{\lambda_2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_1 \underline{s}_1 3^{\lambda_2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_1 \underline{s}_2 3^{\lambda_2} & 0 \end{bmatrix} \Phi + U^* \ .$$

The first $n = 2$ observations center the diagonals of Φ_1 at ones (y_{it} persists at the level of $y_{i,t-1}$ regardless of other variables and further lags). $\lambda_1 \in \mathbb{R}_+$ determines the confidence in this belief. The noisier a series y_{it}, the more we shrink its equation (i.e. $\Phi_{1,ii}$) to the random walk assumption.[34] The second and third set of $n = 2$ observations center the diagonal

[34]For example, the first observation implies $\lambda_1 \underline{s}_1 = \lambda_1 \underline{s}_1 \Phi_{1,11} + u_{11}^*$, which under $u_{11}^* \sim N(0, \Sigma_{11})$ yields

elements of Φ_2 and Φ_3 at zero. For higher lags, we do that with increasing confidence, as determined by $\lambda_2 \in \mathbb{R}_{++}$.

The second block of $n\lambda_3$ observations determines the prior for Σ. We stack the following $n = 2$ observations λ_3 times:

$$\begin{bmatrix} \underline{s}_1 & 0 \\ 0 & \underline{s}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Phi + U^* \ .$$

This centers the prior for Σ_{ii} at $\underline{s}_i^2$ and that of Σ_{ij} at zero, and $\lambda_3 \in \mathbb{Z}$ reflects the confidence in this belief.

The third block of $n + 1$ observations are:

$$\begin{bmatrix} \lambda_4 \underline{y}_1 & 0 \\ 0 & \lambda_4 \underline{y}_2 \\ \lambda_5 \underline{y}_1 & \lambda_5 \underline{y}_2 \end{bmatrix} = \begin{bmatrix} \lambda_4 \underline{y}_1 & 0 & \lambda_4 \underline{y}_1 & 0 & \lambda_4 \underline{y}_1 & 0 & 0 \\ 0 & \lambda_4 \underline{y}_2 & 0 & \lambda_4 \underline{y}_2 & 0 & \lambda_4 \underline{y}_2 & 0 \\ \lambda_5 \underline{y}_1 & \lambda_5 \underline{y}_2 & \lambda_5 \underline{y}_1 & \lambda_5 \underline{y}_2 & \lambda_5 \underline{y}_1 & \lambda_5 \underline{y}_2 & \lambda_5 \end{bmatrix} \Phi + U^* \ .$$

The first $n = 2$ observations reflect the view that if y_{it} was a at a certain level in the past three lags, then we expect it to stay at that level. $\lambda_4 \in \mathbb{R}_+$ determines the confidence in this belief. The last observation reflects the belief that the variables in y_t tend to jointly persist at the same level. $\lambda_5 \in \mathbb{R}_+$ determines the confidence in this belief. This last block of observations determines the prior of Φ_0 and influences the priors of Φ_1, Φ_2, Φ_3 (thire off-diagonal elements mainly).

Posterior of (Φ, Σ) in Reduced-Form VAR Under MNIW-Prior The (conditional) likelihood for a reduced-form VAR is

$$p(Y|\Phi, \Sigma) = (2\pi)^{-nT/2}|\Sigma|^{-T/2} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(Y - X\Phi)'(Y - X\Phi)\right]\right\} \ .$$

As stated in the main text, combined with the MNIW-prior

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

it leads to the MNIW-posterior

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

$\Phi_{1,11} \sim N\left(1, \frac{\Sigma_{11}}{(\lambda_1 \underline{s}_1)^2}\right)$.

with $\bar{P}, \bar{\mu}, \bar{\nu}$ and $\bar{S}$ as stated in the main text. Analogously to the posterior for (β, σ^2) under a Normal likelihood and Normal-Inverse Gamma-prior, this posterior is derived as follows.

First, we derive the conditional posterior of $\Phi|\Sigma$. The prior pdf of $\Phi|\Sigma$ is

$$\begin{aligned} p(\Phi|\Sigma) &= (2\pi)^{-nk/2}|\Sigma|^{-k/2}|\underline{P}^{-1}|^{-n/2} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(\Phi-\underline{\mu})'\underline{P}(\Phi-\underline{\mu})\right]\right\} \\ &\propto exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(\Phi'\underline{P}\Phi - 2\Phi'\underline{P}\underline{\mu})\right]\right\}, \end{aligned}$$

where the second line drops all terms that do not involve Φ and makes use of the fact that $\Phi'\underline{P}\underline{\mu} = \underline{\mu}'\underline{P}\Phi$. In turn, we get

$$\begin{aligned} p(\Phi|Y,\Sigma) &= \frac{p(Y|\Phi,\Sigma)p(\Phi|\Sigma)}{p(Y|\Sigma)} \propto p(Y|\Phi,\Sigma)p(\Phi|\Sigma) \\ &\propto exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(Y-X\Phi)'(Y-X\Phi)\right]\right\} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(\Phi-\underline{\mu})'\underline{P}(\Phi-\underline{\mu})\right]\right\} \\ &\propto exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}(\Phi'(\underline{P}+X'X)\Phi - 2\Phi'(X'Y+\underline{P}\underline{\mu}))\right]\right\}. \end{aligned}$$

Defining $\bar{P} = \underline{P} + X'X$ and $\bar{P}\bar{\mu} = X'Y + \underline{P}\underline{\mu}$ – which yields $\bar{\mu} = \bar{P}^{-1}\left[X'Y + \underline{P}\underline{\mu}\right]$ –, this expression looks analogous to that for the pdf of a Matrix-Normal distribution $MN(\bar{\mu}, \bar{P}^{-1}, \Sigma)$ (see above expression for $p(\Phi|\Sigma)$), from which we conclude that $\Phi|Y,\Sigma \sim MN(\bar{\mu}, \bar{P}^{-1}, \Sigma)$.

Second, we invert Bayes' formula above to find

$$\begin{aligned} p(Y|\Sigma) &= \frac{p(Y|\Phi,\Sigma)p(\Phi|\Sigma)}{p(\Phi|Y,\Sigma)} \\ &= (2\pi)^{-nT/2}|\Sigma|^{-T/2}|\underline{P}|^{n/2}|\bar{P}|^{-n/2} exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1}M\right]\right\}, \end{aligned}$$

where $M = Y'Y + \underline{\mu}'\underline{P}\underline{\mu} - \bar{\mu}'\bar{P}\bar{\mu}$. This expression is obtained by first inserting the full expressions for the pdfs $p(Y|\Phi,\Sigma)$, $p(\Phi|\Sigma)$ and $p(\Phi|Y,\Sigma)$ on the RHS and then cancelling all terms that involve Φ, as $p(Y|\Sigma)$ is not a function of Φ. Also, note that $|\underline{P}^{-1}|^{-n/2} = |\underline{P}|^{n/2}$.

Third, we use $p(Y|\Sigma)$[35] to derive the marginal posterior of Σ. Note that the prior pdf of Σ

[35]This is the marginal likelihood conditional on Σ. In other words, it is the likelihood with Φ integrated out using its conditional prior: $p(Y|\Sigma) = \int p(Y|\Phi,\Sigma)p(\Phi|\Sigma)d\Phi$.

is

$$\begin{aligned} p(\Sigma) &= \frac{|\underline{S}|^{\underline{\nu}/2}}{2^{\underline{\nu}n/2}\Gamma_n(\frac{\underline{\nu}}{2})}|\Sigma|^{-\frac{\underline{\nu}+n+1}{2}} exp\left\{-\frac{1}{2}tr\left[\underline{S}\Sigma^{-1}\right]\right\} \\ &\propto |\Sigma|^{-\frac{\underline{\nu}+n+1}{2}} exp\left\{-\frac{1}{2}tr\left[\underline{S}\Sigma^{-1}\right]\right\} . \end{aligned}$$

For the posterior, we then get

$$\begin{aligned} p(\Sigma|Y) &\propto p(Y|\Sigma)p(\Sigma) \\ &\propto |\Sigma|^{-T/2} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}M\right]\right\}|\Sigma|^{-\frac{\underline{\nu}+n+1}{2}} exp\left\{-\frac{1}{2}tr\left[\underline{S}\Sigma^{-1}\right]\right\} \\ &= |\Sigma|^{-\frac{\underline{\nu}+T+n+1}{2}} exp\left\{-\frac{1}{2}tr\left[\Sigma^{-1}(\underline{S}+M)\right]\right\} , \end{aligned}$$

from which we can deduce that $\Sigma|Y \sim IW(\bar{S}, \bar{\nu})$ with $\bar{\nu} = \underline{\nu} + T$ and $\bar{S} = \underline{S} + M$.

Finally, we obtain the marginal likelihood $p(Y)$[36] by inverting Bayes' theorem above, inserting the expressions for the pdfs on the RHS, and cancelling all terms that depend on Σ:

$$\begin{aligned} p(Y) &= \frac{p(Y|\Sigma)p(\Sigma)}{p(\Sigma|Y)} \\ &= \pi^{-nT/2}|\underline{P}|^{n/2}|\bar{P}|^{-n/2}|\underline{S}|^{\underline{\nu}/2}|\bar{S}|^{-\bar{\nu}/2}\Gamma_n\left(\frac{\bar{\nu}}{2}\right)/\Gamma_n\left(\frac{\underline{\nu}}{2}\right) .^{37} \end{aligned}$$

Likelihood of GARCH$(1,1)$

Recall the GARCH$(1,1)$ process from Eq. (10.14):

$$y_t \sim N(0, h_t) , \quad h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1} .$$

[36] This is the unconditional marginal likelihood, i.e. the likelihood with both Φ as well as Σ integrated out.

[37] Thereby, the expression $2^{-\underline{\nu}n/2}2^{-\bar{\nu}n/2} = 2^{Tn/2}$ cancels with the 2 in the first term, $(2\pi)^{-nT/2}$.

Note that h_t, the variance of y_t, is only a function of the initial variance h_0 and all past observations $Y_{0:t-1}$:

$$\begin{aligned} h_t &= \omega + \alpha y_{t-1}^2 + \beta h_{t-1} \\ &= \omega + +\beta\omega + \alpha y_{t-1}^2 + \beta\alpha y_{t-2}^2 + \beta^2 h_{t-2} \\ &= ... \\ &= \sum_{j=0}^{t-1} \beta^j \omega + \sum_{j=0}^{t-1} \beta^j \alpha y_{t-j-1}^2 + \beta^t h_0 \\ &= \frac{1-\beta^t}{1-\beta}\omega + \alpha \sum_{j=0}^{t} \beta^{j-1} y_{t-j}^2 + \beta^t h_0 \ . \end{aligned}$$

As a result, we can derive the likelihood simply as the product of conditional densities:

$$\begin{aligned} p(Y|y_0, h_0, \theta) &= \prod_{t=1}^{T} p(y_t|Y_{0:t-1}, h_0, \theta) \\ &= \prod_{t=1}^{T} (2\pi h_t)^{-1/2} exp\left\{-\frac{1}{2h_t} y_t^2\right\} \\ &= (2\pi)^{-T/2} \left[\prod_{t=1}^{T} h_t\right]^{-1/2} exp\left\{-\frac{1}{2}\sum_{t=1}^{T} \frac{y_t^2}{h_t}\right\} \ , \end{aligned}$$

where h_t is given by the expression above, and $\theta = (\omega, \alpha, \beta)$ contains all the unknown parameters. We could add h_0 as a parameter to-be-estimated or simply set it to zero.

Further Notes on Smoothing Algorithms

Derivation of Kalman Smoother Given the densities $\{p(s_t|s_{t+1}, Y_{1:t}) = N(s_{t|t+1}, P_{t|t+1})\}_{t=1}^T$ from the CK simulation smoother, we can construct the densities $\{p(s_t|Y_{1:T}) = N(s_{t|T}^*, P_{t|T}^*)\}_{t=1}^T$ for the Kalman smoother sequentially as follows. At $t = T$, note that we have $p(s_T|Y_{1:T}) = N(s_{T|T}, P_{T|T})$ from the Kalman filter. To get $p(s_{T-1}|Y_{1:T}) = N(s_{T-1|T}^*, P_{T-1|T}^*)$, use the density $p(s_{T-1}|s_T, Y_{1:T}) = N(s_{T-1|T}, P_{T-1|T})$ and the LIE:

$$\begin{aligned} s_{T-1|T}^* = \mathbb{E}[s_{T-1}|Y_{1:T}] &= \mathbb{E}\left[\mathbb{E}[s_{T-1}|s_T, Y_{1:T}]\right] \\ &= \mathbb{E}[s_{T-1|T}] \\ &= \mathbb{E}\left[s_{T-1|T-1} + P_{T-1|T-1}\Phi_1' P_{T|T-1}^{-1}(s_T^m - s_{T|T-1})\right] \\ &= s_{T-1|T-1} + P_{T-1|T-1}\Phi_1' P_{T|T-1}^{-1}(s_{T|T} - s_{T|T-1}) \ , \end{aligned}$$

and

$$\begin{aligned}
P^*_{T-1|T} = \mathbb{V}[s_{T-1}|Y_{1:T}] &= \mathbb{V}\left[\mathbb{E}[s_{T-1}|s_T, Y_{1:T}]\right] + \mathbb{E}\left[\mathbb{V}[s_{T-1}|s_T, Y_{1:T}]\right] \\
&= \mathbb{V}[s_{T-1|T}] + \mathbb{E}\left[P_{T-1|T}\right] \\
&= \mathbb{V}\left[s_{T-1|T-1} + P_{T-1|T-1}\Phi_1' P^{-1}_{T|T-1}(s^m_T - s_{T|T-1})\right] + P_{T-1|T} \\
&= P_{T-1|T-1}\Phi_1' P^{-1}_{T|T-1} P_{T|T} P^{-1}_{T|T-1}\Phi_1 P_{T-1|T-1} + P_{T-1|T} \\
&= P_{T-1|T-1} + P_{T-1|T-1}\Phi_1' P^{-1}_{T|T-1}(P_{T|T} - P_{T|T-1})P^{-1}_{T|T-1}\Phi_1 P_{T-1|T-1} \ ,
\end{aligned}$$

Continuing this way for the remaining densities $p(s_t|Y_{1:T}) = N(s^*_{t|T}, P^*_{t|T})$ for $t = T-2, ..., 1$ gives analogous formulas, as written out in the main text.

Smoothing Algorithms for Companion-Form-VAR(1) The CK Simulation Smoother and the Kalman Smoother need to be adjusted for cases in which $s_t|s_{t+1}$ is not fully stochastic, but contains deterministic elements. An instructive example is the companion-form-VAR(1), discussed in Nelson and Kim (1999, p. 194). As explained in Section 10.2, in that case $s_t = (y_t', y_{t-1}', ..., y_{t-p+1}')'$ is an $np \times 1$ vector stacking p times the same $n \times 1$-process y_t at different lags. Denote the first n elements of s_t as $[s_t]_1 = y_t$, denote the $n \times np$-matrix containing the first n rows of the $np \times np$ matrix Φ_1 by $[\Phi_1]_{1\cdot}$, and denote the upper-left $n \times n$ block of Σ as $[\Sigma]_{11}$. The smoothing algorithms have to be adjusted as follows.

Algorithm 27 (CK Simulation Smoother for Companion Form-VAR(1))**.**

1. *Run the Kalman filter to get* $\{s_{t|t}, s_{t|t-1}, P_{t|t}, P_{t|t-1}\}_{t=1}^T$.
2. *Draw* $[s^m_T]_1$ *from* $N\left(\left[s_{T|T}\right]_1, \left[P_{T|T}\right]_{11}\right)$.
3. *For* $t = T-1, ..., 0$*, given draw* $\left[s^m_{t+1}\right]_1$ *from* $N\left(\left[s_{t+1|t+2}\right]_1, \left[P_{t+1|t+2}\right]_{11}\right)$*, draw* $[s^m_t]_1$ *from* $N\left(\left[s_{t|t+1}\right]_1, \left[P_{t|t+1}\right]_{11}\right)$ *with*

   - $s_{t|t+1} = s_{t|t} + P_{t|t}[\Phi_1]_{1\cdot}' H^{-1}(\left[s^m_{t+1}\right]_1 - \left[s_{t+1|t}\right]_1)$,
   - $P_{t|t+1} = P_{t|t} - P_{t|t}[\Phi_1]_{1\cdot}' H^{-1}[\Phi_1]_{1\cdot} P_{t|t}$,

   where $H = [\Phi_1]_{1\cdot} P_{t|t}[\Phi_1]_{1\cdot}' + [\Sigma]_{11}$.

Relative to the notation used for the Kalman filter, this is with a slight abuse of notation, as $s_{t|t+1} \neq \mathbb{E}\left[s_t|X_{1:t}, s_{t+1}\right]$ but $s_{t|t+1} = \mathbb{E}\left[s_t|X_{1:t}, [s_{t+1}]_1\right]$, and similarly $P_{t|t+1} = \mathbb{V}\left[s_t|X_{1:t}, [s_{t+1}]_1\right]$.

Algorithm 28 (Kalman Smoother for Companion Form-VAR(1))**.**

1. *Run the Kalman filter to get* $\{s_{t|t}, s_{t|t-1}, P_{t|t}, P_{t|t-1}\}_{t=1}^{T}$.
2. *We know* $[s_T]_1|Y_{1:T} \sim N([s_{T|T}]_1, [P_{T|T}]_{11})$.
3. *For* $t = T-1, ..., 1$, *given* $[s_{t+1}]_{11}|Y_{1:T} \sim N([s^*_{t+1|T}]_1, [P^*_{t+1|T}]_{11})$, *we get*

$$[s_t]_1|Y_{1:T} \sim N([s^*_{t|T}]_1, [P^*_{t|T}]_{11}) \ , \quad \textit{with} \quad \begin{array}{l} s^*_{t|T} = s_{t|t} + P_{t|t}[\Phi_1]'_{1.}H^{-1}([s^*_{t+1|T}]_1 - [s_{t+1|t}]_1) \\ P^*_{t|T} = P_{t|t} + P_{t|t}[\Phi_1]'_{1.}H^{-1}([P^*_{t+1|T}]_{11} - H)H^{-1}\Phi_1 P_{t|t} \end{array} ,$$

where $H = [\Phi_1]_{1.}P_{t|t}[\Phi_1]'_{1.} + [\Sigma]_{11}$.

DSGE Models as State Space Models

Consider the Real Business Cycle (RBC) model, the predecessor of all modern Dynamic Stochastic General Equilibrium (DSGE) models. It is defined by the optimization problems of a representative firm and a representative household, whereby Total Factor Productivity (TFP) z_t varies exogenously according to an AR(1) process:

$$z_t = \rho z_{t-1} + \varepsilon_t \ , \quad \varepsilon_t \sim N(0, \sigma^2) \ . \tag{10.16}$$

For equilibrium, we require that the firm and household act optimally and that market clearing holds:

- The representative firm maximizes profits by solving

$$\max_{L_t, K_t} y_t - w_t L_t - r_t K_t \ ,$$

  whereby $y_t = e^{z_t} K_t^{\alpha} L_t^{1-\alpha}$ is production, and the firm takes TFP z_t, wages w_t and the rental rate of capital r_t as given.

- The representative household maximizes its lifetime utility. The optimization problem can be represented by the Bellman equation

$$\begin{aligned} V(k_t, z_t, K_t) = \max_{c_t, l_t, k_{t+1}} \ & u(c_t, l_t) + \beta \mathbb{E}_t \left[V(k_{t+1}, z_{t+1}, K_{t+1}) \right] \\ \text{s.t.} \quad & c_t + i_t = w_t l_t + r_t k_t \ , \\ & i_t = k_{t+1} - (1-\delta) k_t \ , \end{aligned}$$

  whereby $u(c, l) = \frac{c^{1-\tau}}{1-\tau} - \frac{l^{1-\kappa}}{1-\kappa}$ is the household's utility and V is its value function.

The expectation is taken w.r.t. z_{t+1} and K_{t+1}, taking z_t and K_t as given (and knowing/having rational expectations w.r.t. the processes of z_{t+1} and K_{t+1}). Note that k_{t+1} is known as it is chosen by the household.

- The household is indeed representative: $k_{t+1} = K_{t+1}$, i.e. the aggregate capital stock K_{t+1} evolves as the capital stock chosen by the representative household, k_{t+1}.
- The labor market and goods market clear:

$$l_t = L_t \ , \quad c_t + i_t = y_t \ ,$$

where $y_t = e^{z_t} k_t^{\alpha} l_t^{1-\alpha}$ is aggregate production.

The firm's first-order conditions (FOCs) imply

$$w_t = (1-\alpha) e^{z_t} k_t^{\alpha} l_t^{-\alpha} \ , \tag{10.17}$$

$$r_t = \alpha e^{z_t} k_t^{\alpha-1} l_t^{1-\alpha} \ . \tag{10.18}$$

To solve the household's problem, solve the budget constraint for c_t and plug it into the utility function to get a maximization problem w.r.t. two variables only, l_t and k_{t+1}. The FOC w.r.t. l_t gives $u_1(c_t, l_t) w_t + u_2(c_t, l_t) = 0$, where u_1 denotes the derivative of u w.r.t. its first argument and analogously for u_2. This gives

$$l_t^{\kappa} = w_t c_t^{-\tau} \ . \tag{10.19}$$

The FOC w.r.t. k_{t+1} gives $-u_1(c_t, l_t) + \beta \mathbb{E}[V_1(k_{t+1}, z_{t+1}, K_{t+1})] = 0$, which together with the fact that $V_1(k_t, z_t, K_t) = u_1(c_t, l_t)(r_t + 1 - \delta)$ gives the so-called Euler equation

$$c_t^{-\tau} = \beta \mathbb{E}_t \left[c_{t+1}^{-\tau} (r_{t+1} + 1 - \delta) \right] \ . \tag{10.20}$$

Eqs. (10.17) to (10.20), together with the market-clearing equation

$$c_t + k_{t+1} - (1-\delta) k_t = e^{z_t} k_t^{\alpha} l_t^{1-\alpha} \ . \tag{10.21}$$

form a system of five equations for five unknowns: $c_t, l_t, k_{t+1}, w_t, r_t$. Given k_t and z_t, we can solve for these five unknowns.[38] We say we solved a dynamic macroeconomic model when we found so-called policy functions for each of these five endogenous variables, i.e. functions that, given values for the two variables k_t and z_t, return the value for these five variables

[38] In fact, these equilibrium conditions are the same at every point in time t. Thus, given an initial k_0 and the whole process $\{z_0\}_{t=1}^T$, we can solve for the sequences of these five unknowns: $\{c_t, l_t, k_{t+1}, w_t, r_t\}_{t=0}^T$.

that solve the equilibrium conditions Eqs. (10.17) to (10.21).[39] We can write these policy functions compactly as a single, vector-valued function $g(z, k)$, which is nonlinear because the equilibrium conditions are nonlinear. Once we have it, we can plug in k_t and z_t at any t to get $(c_t, l_t, k_{t+1}, w_t, r_t)' = g(z_t, k_t)$. Together with the AR(1) equation for the exogenous process z_t, this defines a nonlinear process for the vector $s_t = (c_t, l_t, k_{t+1}, w_t, r_t, z_t)'$! Given $s_{t-1} = (c_{t-1}, l_{t-1}, k_t, w_{t-1}, r_{t-1}, z_{t-1})'$ as well as the realization of the shock ε_t, the function g and the AR(1)-equation for z_t tell us how we get s_t:

$$s_t = \Phi(s_{t-1}, \varepsilon_t; \theta) \ . \tag{10.22}$$

Thereby, $\theta = \{\alpha, \rho, \sigma^2, \beta, \tau, \kappa, \delta\}$ contains all unknown parameters. To estimate it, we use data on some of the variables in s_t or functions thereof. Denote this data as y_t and write

$$y_t = \Psi(s_t; \theta) + \eta_t \ , \tag{10.23}$$

where η_t is a measurement error, which is sometimes included to acknowledge that the model might have difficulty in accounting for the data by itself.

Eqs. (10.22) and (10.23) define a nonlinear SS model! We can use it to estimate the unknown parameters θ, which allows us to analyze properties of the model while acknowledging our uncertainty about the exact values of θ as well as to see how well the model fits the data $\{y_t\}_{t=1}^T$. While typically the function Ψ is known to the researcher and oftentimes even linear, the function Φ is not always easy to find and is nonlinear, owing to the fact that the model's equilibrium conditions are nonlinear.

There are two classes of methods to solve a macroeconomic model, i.e. to find the policy function g and therefore the function Φ that appears in the transition equation Eq. (10.22). The first approach models g with some functional approximation method and searches numerically for the g that comes closest to solving the equilibrium conditions. Examples are grid-based methods like value- or policy-function iteration, projection methods like Chebyshev polynomials or machine-learning methods like neural networks. The other approach, the so-called perturbation methods, involve linearizing the equilibrium conditions around the steady state. The first-order linearization gives a set of linear equations (i.e. linear in variables s_t), which are solved by a linear policy function g. The second-order linearizations gives a set of equations with at most quadratic terms in s_t, which is solved by a policy function g that has at most quadratic terms in z_t and k_t. This pattern continues for higher-order lin-

[39] In macroeconomics, k_t and z_t are called state variables and the variables $(c_t, l_t, k_{t+1}, w_t, r_t)$ are called policy variables. This differs from the econometric definition of a state variable, which is why it is left out from the main text so that no confusion arises.

earizations. Perturbation methods are only useful for analyzing dynamics that evolve tightly around the steady state under models that are close-to-linear; they generate distorted dynamics in cases where the variables in the model significantly depart from the steady state and/or the model contains major nonlinearities like asymmetric adjustment costs, stochastic volatility or ocassionally-binding constraints (e.g. the zero-lower bound on interest rates or a borrowing constraint). The bottom line is that all solution methods but the first-order linearization give rise to a nonlinear SS model, whereby the function Φ is not even known analytically, but only numerically, i.e. given s_{t-1} and ε_t, we numerically obtain a value for s_t.

In the following, we solve the above macroeconomic model using a first-order linearization around the steady state and we derive the linear SS model implied by it. For this purpose, we first find the steady state of the model, i.e. the set of variables $s = (c, l, k, w, r, z)$ that solve the system of equations when there are no shocks: $\varepsilon = 0$. Trivially, by Eq. (10.16), we have $z = 0$. Eq. (10.20) becomes $c^{-\tau} = \beta\left[c^{-\tau}(r + 1 - \delta)\right]$ in steady state, which implies

$$r = \beta^{-1} - 1 + \delta \ .^{40}$$

Solving Eq. (10.17) and Eq. (10.18) for k/l and equating the two expressions yields

$$w = (1-\alpha)^{\frac{1}{\alpha}}(\alpha/r)^{\frac{1}{1-\alpha}} \ .$$

Solving for k, c and l is a bit more involved. By Eq. (10.18), we get $k = (\alpha/r)^{\frac{1}{1-\alpha}} l$. Plugging in this expression in Eq. (10.21) yields $c = \gamma l$, with γ as defined below. Plugging this expression for c as well as the above expression for w into Eq. (10.19) yields

$$l = \left[(1-\alpha)^{\frac{1}{\alpha}}(\alpha/r)^{\frac{1}{1-\alpha}}\gamma^{-\tau}\right]^{\frac{1}{\tau+\kappa}} \ , \quad \text{with} \quad \gamma = \left[(\alpha/r)^{\frac{\alpha}{1-\alpha}} - \delta(\alpha/r)^{\frac{1}{1-\alpha}}\right] \ .$$

In turn, this implies

$$c = \left[(1-\alpha)^{\frac{1}{\alpha}}(\alpha/r)^{\frac{1}{1-\alpha}}\gamma^{\kappa}\right]^{\frac{1}{\tau+\kappa}} \quad \text{and} \quad k = \left[(1-\alpha)^{\frac{1}{\alpha}}(\alpha/r)^{\frac{\tau+\kappa+1}{1-\alpha}}\gamma^{-\tau}\right]^{\frac{1}{\tau+\kappa}} \ .$$

Next, we linearize the equilibrium conditions Eqs. (10.17) to (10.21) around this steady state.

[40] $\beta \in (0,1)$ is the discount factor of the household; it captures how much the household values next period's consumption relative to today's consumption. We can write it also as $\beta = \frac{1}{1+d}$, where $d > 0$ is the household's discount rate. The above equation says that, in steady state, the rental rate of capital, i.e. the return that the household obtains on its savings, must be equal to the discount rate $d = \beta^{-1} - 1$ plus the depreciation rate δ. In other words, after subtracting depreciation of the capital stock, the household must be left with a "net return" equal to the discount rate. Only then can it be induced to have a flat profile of consumption.

For Eq. (10.17), we obtain

$$\begin{aligned}(w_t - w) =&(1-\alpha)e^z k^\alpha l^{1-\alpha}(z_t - z)\\ &+(1-\alpha)e^z l^{1-\alpha}\alpha k^{\alpha-1}(k_t - k)\\ &+(1-\alpha)e^z k^\alpha(-\alpha)l^{-\alpha}(l_t - l)\ .\end{aligned}$$

Note that $w = (1-\alpha)e^z k^\alpha l^{1-\alpha}$ as well as $z = 0$ holds in steady state. This gives us an equation that relates the absolute deviations of w_t, z_t, k_t and l_t from their respective steady state values:

$$(w_t - w) = wz_t + \alpha w k^{-1}(k_t - k) + -\alpha w l^{-1}(l_t - l)\ .$$

Dividing both sides by w, we obtain an equation that relates the percentage devations of w_t, k_t and l_t as well as the absolute deviation of z_t from their respective steady state values:

$$\hat{w}_t = z_t + \alpha\hat{k}_t - \alpha\hat{l}_t\ . \tag{10.24}$$

Repeating the same steps for Eq. (10.18) gives

$$\hat{r}_t = z_t + (\alpha-1)\hat{k}_t + (1-\alpha)\hat{l}_t\ , \tag{10.25}$$

while for Eq. (10.19) we get

$$\kappa\hat{l}_t = \hat{w}_t - \tau\hat{c}_t\ . \tag{10.26}$$

Linearizing equations with additive terms leads to less appealing equations. For Eq. (10.21), we obtain

$$\frac{c}{y}\hat{c}_t + \frac{k}{y}\hat{k}_{t+1} - (1-\delta)\frac{k}{y}\hat{k}_t = z_t + \alpha\hat{k}_t + (1-\alpha)\hat{l}_t\ . \tag{10.27}$$

Finally, we linearize Eq. (10.20), keeping the (conditional) expectation operator:

$$-\tau\hat{c}_t = -\tau\mathbb{E}_t[\hat{c}_{t+1}] + \lambda\mathbb{E}_t[\hat{r}_{t+1}]\ , \quad \text{with} \quad \lambda = (1-\beta(1+\delta))\ .$$

In the subsequent derivations, special attention is given to the expectation operators that appear in these equilibrium conditions. For this purpose, it is useful to rewrite $\hat{r}_{t+1}$ in terms of $\hat{c}_{t+1}$, z_{t+1} and $\hat{k}_{t+1}$ because this leaves us with a single "forward-looking variable" $\hat{c}_{t+1}$, while we know $\hat{k}_{t+1}$ and we can easily compute the conditional expectation of z_{t+1}: $\mathbb{E}_t[z_{t+1}] = \rho z_t$. Combining equations Eqs. (10.24) to (10.26) so as to cancel $\hat{l}_t$ and $\hat{w}_t$ gives

us

$$\hat{r}_t = \frac{1+\kappa}{\alpha+\kappa} z_t - (1-\alpha)\frac{\kappa}{\alpha+\kappa}\hat{k}_t - (1-\alpha)\tau\frac{1}{\alpha+\kappa}\hat{c}_t \ .$$

In turn, plugging the corresponding expression for $\hat{r}_{t+1}$ into the equation above yields

$$-\tau\hat{c}_t = -\tau\left(1-\lambda\frac{1-\alpha}{\alpha+\kappa}\right)\mathbb{E}_t[\hat{c}_{t+1}] + \lambda\frac{1+\kappa}{\alpha+\kappa}\mathbb{E}_t[z_{t+1}] - \lambda(1-\alpha)\frac{\kappa}{\alpha+\kappa}\hat{k}_{t+1} \ . \tag{10.28}$$

The five equilibrium conditions in Eqs. (10.24) to (10.28) together with the AR(1) equation for the exogenous process z_t in Eq. (10.16) define a linear process for the six-dimensional vector $(\hat{c}_t, \hat{l}_t, \hat{k}_{t+1}, \hat{w}_t, \hat{r}_t, z_t)$. However, deriving an expression for this linear process is challenging due to the presence of forward-looking terms, $\mathbb{E}_t[\hat{c}_{t+1}]$ in this particular model. While we know $\mathbb{E}_t[z_{t+1}] = \rho z_t$, $\mathbb{E}_t[\hat{c}_{t+1}]$ could in principle be anything, and depending on the expectations of the agents, we might get a stable (non-explosive) or an explosive process. Typically, one focuses on the stable solution obtained under rational expectations. The approach of Sims (2001) is to include $\mathbb{E}_t[\hat{c}_{t+1}]$ as a state variable in s_t and augment the system of equations with rational expectations errors $\eta_t = c_t - \mathbb{E}_{t-1}[\hat{c}_t]$. As a result, for $s_t = (\hat{c}_t, \hat{l}_t, \hat{k}_{t+1}, \hat{w}_t, \hat{r}_t, z_t, \mathbb{E}_t[\hat{c}_{t+1}])'$, we can write

$$\Gamma_0(\theta)s_t = \Gamma_1(\theta)s_{t-1} + \Psi(\theta)e_t + \Pi\eta_t \ , \ \ e_t \sim N(0,1) \ . \tag{10.29}$$

Let s_t be $n \times 1$, where $n = 7$ in our particular model. A generalized complex Schur (QZ) decomposition of Γ_0 and Γ_1 yields the $n \times n$ matrices Q, Z, Λ and Ω solving

$$Q'\Lambda Z' = \Gamma_0 \ , \ \ Q'\Omega Z' = \Gamma_1 \ , \ \ QQ' = ZZ' = I \ ,$$

where Λ and Ω are upper-triangular. Defining $w_t = Z's_t$ and premultiplying equation (10.29) by Q yields

$$\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ 0 & \Lambda_{22} \end{bmatrix} \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ 0 & \Omega_{22} \end{bmatrix} \begin{bmatrix} w_{1,t-1} \\ w_{2,t-1} \end{bmatrix} + \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} (\Psi e_t + \Pi\eta_t) \ , \tag{10.30}$$

where it is assumed, without loss of generality, that s_t is ordered in such a way that the $m \times 1$ vector $w_{2,t}$ is explosive, $0 \leq m \leq n$.[41] A non-explosive solution for s_t exists if $w_{2,0} = 0$ and

[41] Concretely, following Sims (2001), the system is ordered such that the set of generalized Eigenvalues (the ratios of the diagonal elements in Ω and Λ: ω_{ll}/λ_{ll}) is such that $|\omega_{ll}/\lambda_{ll}| < \bar{\xi}$ for all $l \leq m$ and $|\omega_{ll}/\lambda_{ll}| \geq \bar{\xi}$ for all $l > m$. This means that the m Eigenvalues largest in absolute size are ordered last. It's crucial to notice that the Schur decomposition is not unique, while the generalized Eigenvalues (usually) are. Thus, one has to specify an ordering first and then compute the corresponding decomposition. This is not the same as ordering the matrices after the decomposition has been completed. This can be seen from the fact that in the latter case the matrices Ω and Λ will generally not be upper-triangular.

for every vector of structural shocks ε_t, one can find a $k \times 1$ vector of rational expectations errors η_t which offset the impact of ε_t on $w_{2,t}$:

$$Q_2 \Psi e_t + Q_2 \Pi \eta_t = 0 \ .$$

In this case, $w_{2,t} = 0 \ \forall \ t$. Sims (2001) notes that a necessary and sufficient condition for the non-explosive solution to be unique is that the row space of $Q_1\Pi$ is contained in that of $Q_2\Pi$, in which case we can write

$$Q_1 \Pi = \Phi Q_2 \Pi \ ,$$

for some matrix Φ (which can be solved for; Φ is $(n-m) \times m$). One can premultiply equation (10.30) by $\begin{bmatrix} I & -\Phi \end{bmatrix}$ to get

$$\begin{aligned}
&\begin{bmatrix} \Lambda_{11} & \Lambda_{12} - \Phi\Lambda_{22} \\ 0 & I \end{bmatrix} \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{12} - \Phi\Omega_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_{1,t-1} \\ w_{2,t-1} \end{bmatrix} + \begin{bmatrix} Q_1 - \Phi Q_2 \\ 0 \end{bmatrix} (\Psi \varepsilon_t + \Pi \eta_t) \\
\Rightarrow &\begin{bmatrix} \Lambda_{11} & \Lambda_{12} - \Phi\Lambda_{22} \\ 0 & I \end{bmatrix} \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{12} - \Phi\Omega_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_{1,t-1} \\ w_{2,t-1} \end{bmatrix} + \begin{bmatrix} Q_1 - \Phi Q_2 \\ 0 \end{bmatrix} \Psi e_t \\
\Rightarrow &\begin{bmatrix} \Lambda_{11} & \Lambda_{12} - \Phi\Lambda_{22} \\ 0 & I \end{bmatrix} Z' s_t = \begin{bmatrix} \Omega_{11} & \Omega_{12} - \Phi\Omega_{22} \\ 0 & 0 \end{bmatrix} Z' s_{t-1} + \begin{bmatrix} Q_1 - \Phi Q_2 \\ 0 \end{bmatrix} \Psi e_t \ , \\
&\Rightarrow A s_t = B s_{t-1} + C e_t \ ,
\end{aligned}$$

where the last m equations in this system simply state that $w_{2,t} = 0 \ \forall \ t$. This can be solved for s_t to yield the transition equation for the linear SS representation:

$$\begin{aligned}
s_t &= A^{-1} B s_{t-1} + A^{-1} C e_t \\
&= \Phi_1(\theta) s_{t-1} + \Phi_\varepsilon(\theta) e_t \\
&= \Phi_1(\theta) s_{t-1} + u_t \ , \quad u_t \sim N(0, \Sigma) \ ,
\end{aligned}$$

with $\Sigma = \Phi_\varepsilon(\theta)\Phi_\varepsilon(\theta)'$. Together with a (linear) measurement equation, this defines a linear SS model.