Chapter 6: Likelihood methods

Maximum likelihood (ML) techniques have enjoyed a remarkable come back in the last few years, probably as a consequence of the development of faster computer technology and of the substantial improvement in the specification of structural models. In fact, complex stochastic general equilibrium models have been recently estimated and tested against the data. This represents a shift of attitude relative to the 1980's or the beginning of the 1990's where GMM and related techniques dominated the scene. As we have seen maximum likelihood is a special case of GMM when the scores of the likelihood are used as orthogonality conditions. Nevertheless, (full information) ML differs from GMM in several respects.

In both cases, a researcher starts from a fully specified dynamic stochastic general equilibrium model. However, while with GMM the first order conditions of the maximization are sufficient for estimation and testing, with maximum likelihood the final form, expressing the endogenous variables of the model as a function of the exogenous variables and of the parameters, is needed. As we have seen in Chapter 2, this is not a small enterprise in general and approximations are often needed, transforming nonlinear specifications into a linear ones. The presence of nonlinearities, on the other hand, does not present particular problems for GMM estimation and testing. Moreover, while with GMM one uses only the (limited) information contained in a subset of the equilibrium conditions, e.g. the Euler equations, once the final form is calculated, all the implications of the model must necessarily be taken into account for estimation. Therefore, while with the former one can estimate and test assuming that only some of the equations of the model appropriately characterize the data generating process, such an assumption is untenable when ML is used. An interesting conundrum arises when misspecification is present. Following White (1982), one can show that a quasi-ML estimator of the parameters, obtained when the distribution of the errors is misspecified, has the same asymptotic properties as the correct ML estimator under a set of regularity conditions. However, as we will argue in chapter 7, the misspecification present in DSGE models is unlikely to be reducible to the distributions of the errors. Hence, it is unknown what kind of properties ML estimates have in these setups and care must be used in reporting and interpreting estimates and tests.

With both ML and GMM the final scope of the analysis is the evaluation of the quality of the model's approximation to the data and, given estimates, to study the effects of altering interesting economic (policy) parameters. This should be contrasted with the exercises typically performed in VARs. Here the full implications of the model, as opposed to a set of

minimal restrictions, are used to obtain estimates of the objects of interest; the analysis is geared towards the estimation of "structural parameters' as opposed to "structural shocks"; and model evaluation is often more important than describing the (restricted) structure of the data in response to disturbances. Which approach one subscribes depends on how much a researcher trusts the model. With ML (and GMM) one puts a lot of faith in the model as a description of the data - the structure is correct, only the parameters are unknown. With VARs the opposite is true. Therefore only a limited set of conventional or generic restrictions are considered.

This chapter describes the steps needed to estimate models with ML. We start by describing the use of the Kalman filter and of the Kalman smoother for state space models. State space models are general structures: any multivariate ARMA model and almost all log-linearized DSGE model can be fit into this framework. The Kalman filter, besides providing minimum MSE forecasts of the endogenous variables and optimal recursive estimates of the unobserved states, is an important building block in the prediction error decomposition of the likelihood. In fact, the likelihood function of a state space model can be conveniently expressed in terms of the one-step ahead forecast errors, conditional on the initial observations, and of their recursive variance, both of which can be obtained with the Kalman filter. Therefore, given some initial parameter values, the Kalman filter can be used to recursively construct the likelihood function; gradient methods can be employed to provide new estimates for the parameters and the two-step procedure can be repeated until the gradient or the parameters do not change across iterations.

In the third section we provide some numerical tips on how to update parameter estimates and on other issues often encountered in practice. The algorithms are only sketched here. For details the reader should consult Press et al. (1980) or Judge, et. al (1985). The last portion of this chapter applies the machinery we have developed to the problem of estimating DSGE models. The (log)-linearized solution of such models naturally comes into a state space format where the coefficients are highly nonlinear functions of the structural parameters. We discuss a number of peculiarities of DSGE models relative to other time series specifications and describe how to use cross-equations restrictions to identify structural parameters and to test the model. This is the approach popularized by Sargent (1979) and Sargent and Hansen (1980) and exploits the fact that linearized expectational equations impose restrictions on the VAR of the data. We conclude estimating the parameters of a simple sticky price model driven by technology and monetary disturbances and confronting some of the implied unconditional moments to the data.

6.1 The Kalman filter

The Kalman filter is one of the most important instruments in the toolkit of applied macroeconomists and we will extensively use it throughout the rest of this book. The presentation here is basic and the reader should refer to Harvey (1991) or Anderson and Moore (1979) for more extensive details.

The Kalman filter is typically employed in state space models of the form

$$y_t = x'_{1t}\alpha_t + x'_{2t}v_{1t} \tag{6.1}$$
$$\alpha_t = \mathbb{D}_{0t} + \mathbb{D}_{1t}\alpha_{t-1} + \mathbb{D}_{2t}v_{2t} \tag{6.2}$$

where x'_{1t} is $m \times m_1$ matrix, x'_{2t} is $m \times m_2$ matrix , $\mathbb{D}_{0t}$ is $m_1 \times 1$ vector, $\mathbb{D}_{1t}$, $\mathbb{D}_{2t}$ are $m_1 \times m_1$ and $m_1 \times m_3$ matrices; v_{1t} is a $m_2 \times 1$ vector of martingale difference sequences, $v_{1t} \sim \mathbb{N}(0, \Sigma_{v_1})$; v_{2t} is $m_3 \times 1$ vector of martingale difference sequences, $v_{2t} \sim \mathbb{N}(0, \Sigma_{v_2})$. We also assume that $E(v_{1t}v'_{2\tau}) = 0$ and $E(v_{1t}\alpha'_0) = 0$, for all t and τ. The first assumption can be dispensed of, as we will see later on. The two together insure that the states α_t and the disturbances v_{1t} are uncorrelated.

(6.1) is typically referred as the measurement (observation) equation while (6.2) is the transition (state) equation. Note that, in principle, α_t is allowed to vary over time and that $x_{1t}, x_{2t}, \mathbb{D}_{0t}, \mathbb{D}_{1t}, \mathbb{D}_{2t}$ could be fixed (i.e. matrices of numbers) or realizations of random variables. For example, in time series context x_{1t} could contain lagged y_t's and x_{2t} current and/or lagged stochastic volatility terms. Notice that it is possible to have m_2 shocks driving the m endogenous variables, $m_2 \leq m$.

The framework provided by (6.1)-(6.2) is general: a number of time series and regression models can be cast in such a format. We consider a few special cases next.

Example 6.1 *Consider an m variable VAR $y_t = A(\ell)y_{t-1} + e_t$, where $A(\ell)$ is a polynomial of order q and e_t is a martingale difference process, $e_t \sim (0, \Sigma_e)$. As we have seen such a system can be rewritten in a companion form as $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$ where $\mathbb{A} = [\mathbb{A}_1, \mathbb{A}_2]'$ and $\mathbb{A}_1 = (A_1, \ldots, A_q)'$ contains the first m rows of $\mathbb{A}$, $\mathbb{A}_2$ is a matrix of ones and zeros and $E_t = (e_t, 0, \ldots, 0)'$. Such a system fits into (6.1)-(6.2) setting $\alpha_t = \mathbb{Y}_t = [y'_t, y'_{t-1}, \ldots y'_{t-q}]'$, $x'_{1t} = [I, 0, \ldots 0]$, $\mathbb{D}_{1t} = \mathbb{A}, \Sigma_{v_1} = 0$, $v_{2t} = E_t$, $\mathbb{D}_{2t} = I$, $\mathbb{D}_{0t} = 0$. Hence, there is no measurement error, the measurement equation is trivial and states and observables coincide.*

Example 6.2 *Consider the univariate process, $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t + D_1 e_{t-1}$. This model can be equivalently written as:*

$$y_t = [1\ 0] \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$$

$$\begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ A_2 y_{t-2} + D_1 e_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ D_1 \end{bmatrix} e_t$$

Hence, an ARMA(2,1) structure fits (6.1)-(6.2) setting $\alpha_t = \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$, $\mathbb{D}_{1t} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix}$, $\mathbb{D}_{2t} = \begin{bmatrix} 1 \\ D_1 \end{bmatrix}$, $\mathbb{D}_{0t} = 0$, $x'_{1t} = [1,\ 0], \Sigma_{v_1} = 0, \Sigma_{v_2} = \sigma_e^2$.

Exercise 6.1 *Consider a process of the form $y_{1t} = A_1(\ell)y_{1t-1} + D(\ell)e_t + A_2 y_{2t}$ where y_{2t} represents exogenous variables, $A_1(\ell)$ is of order q_1 and $D(\ell)$ of order q_2. Show the form of the state space model in this case. Display $\mathbb{D}_{1t}, \mathbb{D}_{2t}, x'_{1t}, x'_{2t}$.*

Besides time series models, several structures naturally fit into a state space framework.

Example 6.3 *1) In many economic problems the ex-ante real rate is needed but only the ex-post real rate of interest is computable. In this case we could set* $\alpha_t \equiv r_t^e = i_t - \pi_t^e$ *where* π_t^e *is the expected inflation rate and assume e.g.,* $\alpha_t = \mathbb{D}_1\alpha_{t-1} + v_{2t}$*. The observed real rate then is* $y_t \equiv i_t - \pi_t = \alpha_t + v_{1t}$ *where* v_{1t} *is a measurement error.*
2) A RBC model driven by unit root technology shocks implies that all endogenous variables have a common trend (see King, Plosser, Stock and Watson (1991)). Here $\alpha_t = \alpha_{t-1} + v_{2t}$ *is a one dimensional process;* $x'_{1t} = x'_1$ *are the loadings on the trend and* $x'_{2t} = x'_2$ *are the loadings on everything else (cycle, irregular, etc.).*

Exercise 6.2 *When agents are risk neutral, uncovered interest parity implies that interest rates differentials should be related to the expected change in the exchange rate (see example 2.3 of chapter 5). Cast such a relationship into a state space format carefully defining the matrices* $x'_{1t}, x'_{2t}, \mathbb{D}_{0t}, \mathbb{D}_{1t}, \mathbb{D}_{2t}$*.*

Exercise 6.3 *(Nonlinear state space model) Consider the model* $y_t = \alpha_t + v_{1t},\ \alpha_{t+1} = \alpha_t\theta + v_{2t}$ *and suppose one is interested in* θ*, which is unobservable, as is* α_t*. (In a trend-cycle decomposition,* θ *represents, e.g., the persistence of the trend). Cast the problem in a state space format; show the state vector and display the matrices of the model.*

The Kalman filter can be used to optimally estimate the unobservable state vector α_t and to update estimates when a new observation becomes available. As a byproduct, it also produces recursive forecasts of y_t, consistent with the information available at t.

Suppose we want to compute $\alpha_{t|t}$, the optimal (MSE) estimator of α_t using information up to t; and $\Omega_{t|t}$ the MSE matrix of the forecast errors in the state equation. At this stage we let $x'_{1t} = x'_1,\ x'_{2t} = x'_2, \mathbb{D}_{1t} = \mathbb{D}_1,\ \mathbb{D}_{0t} = \mathbb{D}_0,\ \mathbb{D}_{2t} = \mathbb{D}_2$ be known. We also assume that a sample $\{y_t\}_{t=1}^T$ is available. The Kalman filter algorithm has five steps.

Algorithm 6.1

1) *Select initial conditions. If all eigenvalues of* $\mathbb{D}_1$ *are less then one in absolute value, set* $\alpha_{1|0} = E(\alpha_1)$ *and* $\Omega_{1|0} = \mathbb{D}_1\Omega_{1|0}\mathbb{D}'_1 + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}'_2$ *or* $vec(\Omega_{1|0}) = (I - (\mathbb{D}_1 \otimes \mathbb{D}'_1)^{-1})vec(\mathbb{D}_2\Sigma_{v_2}\mathbb{D}'_2)$*, in which case the initial conditions are the unconditional mean and variance of the process. When some of the eigenvalues of* $\mathbb{D}_1$ *are greater than one, initial conditions cannot be drawn from the unconditional distribution and one needs a guess (say,* $\alpha_{1|0} = 0$*,* $\Omega_{1|0} = \kappa * I$*,* κ *large) to start the iterations.*

2) *Predict* y_t *and construct the mean square of the forecasts using* $t-1$ *information*

$$\begin{aligned} E(y_{t|t-1}) &= x'_1\alpha_{t|t-1} \qquad (6.3) \\ E(y_t - y_{t|t-1})(y_t - y_{t|t-1})' &= E(x'_1(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'x_1) + x'_2\Sigma_{v_1}x_2 \\ &= x'_1\Omega_{t|t-1}x_1 + x'_2\Sigma_{v_1}x_2 \equiv \Sigma_{t|t-1} \qquad (6.4) \end{aligned}$$

3) *Update state equation estimates (after observing y_t):*

$$\alpha_{t|t} = \alpha_{t|t-1} + \Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}(y_t - x_1'\alpha_{t|t-1}) \quad (6.5)$$

$$\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}x_1'\Omega_{t|t-1} \quad (6.6)$$

where $\Sigma_{t|t-1}^{-1}$ is defined in (6.4).

4) *Predict the state equation random variables next period:*

$$\alpha_{t+1|t} = \mathbb{D}_1\alpha_{t|t} + \mathbb{D}_0 = \mathbb{D}_1\alpha_{t|t-1} + \mathbb{D}_0 + \mathsf{K}_t\epsilon_t \quad (6.7)$$

$$\Omega_{t+1|t} = \mathbb{D}_1\Omega_{t|t}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' \quad (6.8)$$

where $\epsilon_t = y_t - x_1'\alpha_{t|t-1}$ is the one-step ahead forecast error in predicting y_t, and $\mathsf{K}_t = \mathbb{D}_1\Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}$ is the Kalman gain.

5) *Repeat steps 2)-4) until $t = T$.*

Note that in step 3) $\Omega_{t|t-1}x_1 = E(\alpha_t - \alpha_{t|t-1})(y_t - x_1'\alpha_{t|t-1})'$. Hence, updated estimates of α_t are computed using the least square projection of $\alpha_t - \alpha_{t|t-1}$ on $y_t - y_{t|t-1}$, multiplied by the prediction error. Similarly, $\Omega_{t|t-1} = E(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'$ is updated using a quadratic form involving the covariance between forecast errors in the two equations and the MSE of the forecasts. Note also that equations (6.7)-(6.8) provide the inputs for the next step of the recursion.

Example 6.4 *Consider extracting a signal α_t, for example, the long run trend of output, given that $\alpha_t = \alpha_{t-1}$ and that the trend is linked to output via $y_t = \alpha_t + v_{1t}$ where v_{1t} is a normal martingale difference process with variance $\sigma_{v_1}^2$. Using (6.6) we have that $\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}(\Omega_{t|t-1} + \sigma_{v_1}^2)^{-1}\Omega_{t|t-1} = \frac{\Omega_{t|t-1}}{1+\frac{\Omega_{t|t-1}}{\sigma_{v_1}^2}} = \frac{\Omega_{t-1|t-1}}{1+\frac{\Omega_{t-1|t-1}}{\sigma_{v_1}^2}}$. Hence, starting from some $\Omega_0 = \bar{\Omega}_0$, we have $\Omega_{1|1} = \frac{\bar{\Omega}_0}{1+\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$; $\Omega_{2|2} = \frac{\bar{\Omega}_0}{1+2\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$; ...; $\Omega_{T|T} = \frac{\bar{\Omega}_0}{1+T\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}$. From (6.5) and (6.7), $\alpha_{T+1|T+1} = \alpha_{T|T} + \frac{\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}{1+T\frac{\bar{\Omega}_0}{\sigma_{v_1}^2}}(y_{T+1} - \alpha_{T|T})$. Hence, as $T \to \infty$, $\alpha_{T+1|T+1} = \alpha_{T|T}$ so that, asymptotically, the contribution of additional observations is negligible.*

Exercise 6.4 *Consider a vector MA process $y_t = e_t + e_{t-1}$ where $e_t \sim \mathbb{N}(0, I)$. Show that the optimal one-step ahead predictor for y_{t+1} is $y_{t+1|t} = \frac{t+1}{t+2}[y_t - y_{t|t-1}]$. Conclude that as $T \to \infty$, the optimal one-step ahead predictor is just last period's forecast error. (Hint: Cast the process into a state space format and apply the Kalman filter).*

Exercise 6.5 *Consider the process $y_t = A_1y_{t-1} + A_2y_{t-2} + e_t$. Here $\alpha_t = [y_t', y_{t-1}']'$, $v_{2t} = [e_t, 0]$, $\mathbb{D}_1 = \begin{bmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{bmatrix}$, $\Sigma_{v_2} = \begin{bmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbb{D}_{0t} = v_{1t} = 0$, $x_1' = [1, 0]$. Show how to start the Kalman filter recursions; compute prediction and updated estimates of α_t for the first two observations.*

Exercise 6.6 *Suppose $y_{1t} = A_t y_{1t-1} + D_t y_{2t} + v_{1t}$ and $\alpha_t = (A_t, D_t) = \alpha_{t-1} + v_{2t}$, where y_{2t} are exogenous variables. Show the updating and prediction equations in this case. How would you handle the case of serially correlated v_{2t}?*

At times, it may be useful to construct estimates of the state vector which, at each t, contains information present in the entire sample. This is the case, in particular, in signal extraction problems; for example, when α_t is a common trend for a vector of y_t, we want estimates at each t to contain all the information available up to T. In this case the Kalman filter can be applied starting from the last observation, working backward through the sample, $t = T-1, \ldots, 1$, using $\alpha_{T|T}, \Omega_{T|T}$ and as initial conditions. That is:

$$\alpha_{t|T} = \alpha_{t|t} + (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\alpha_{t+1|T} - \mathbb{D}_1\alpha_{t|t}) \tag{6.9}$$

$$\Omega_{t|T} = \Omega_{t|t} - (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\Omega_{t+1|T} - \Omega_{t+1|t})(\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})' \tag{6.10}$$

Equations (6.9)-(6.10) define the recursions of the so-called Kalman smoother.

Example 6.5 *Continuing with example 6.4, take $\alpha_{T|T}$ and $\Omega_{T|T}$ as initial conditions. Then $\Omega_{1|t} = \frac{\Omega_{T|T}}{1+T\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}$ and $\alpha_{t|T} = \alpha_{t+1|T} + \frac{\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}{1+T\frac{\Omega_{T|T}}{\sigma_{v_1}^2}}(y_{t|T} - \alpha_{t+1|T})$. Can you guess what $\alpha_{1|T}$ is?*

As a byproduct of the estimation, the Kalman filter allows us to transform (6.1)-(6.2) into a system driven by innovations in the measurement equation. In fact, using (6.5)-(6.7), it is immediate to see that (6.1) and (6.2) are equivalent to

$$y_t = x_{1t}'\alpha_{t|t-1} + \epsilon_t \tag{6.11}$$

$$\alpha_{t+1|t} = \mathbb{D}_1\alpha_{t|t} + \mathbb{D}_0 + \mathsf{K}_t\epsilon_t \tag{6.12}$$

where ϵ_t is the forecast error and $E_t(\epsilon_t\epsilon_t') \equiv \Sigma_{t|t-1}$. Hence, if the Kalman gain K_{t-1} is available and given $(\alpha_{1|0}, \Sigma_{1|0})$, $\alpha_{t|t-1}$ and ϵ_t can be computed recursively at any t. In turn, the Kalman gain is immediately obtained when $\Omega_{t-1|t-1}$ is available.

Exercise 6.7 *The reparametrization in (6.12)-(6.11) is trivial in the case of a constant coefficient VAR(q), since it is always possible to rewrite the measurement equation as $y_t = E[y_t|\mathcal{F}_{t-1}] + \epsilon_t$, where $\mathcal{F}_{t-1}$ is the information set at $t-1$. Show how to transform the ARMA(2,1) model of example 6.2 to fit such a representation.*

Hansen and Sargent (1998, pp.126-128) show that equation (6.6) can also be written as $\Omega_{t|t} = \mathbb{D}_1\Omega_{t-1|t-1}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' - \mathbb{D}_1\Omega_{t-1|t-1}x_1\Sigma_{t|t-1}^{-1}x_1'\Omega_{t-1|t-1}\mathbb{D}_1$. One can recognize in this expression a version of the matrix Riccati equation used in chapter 2 to solve linear regulator problems. Therefore, under regularity conditions, in state space models with constant coefficients, $\lim_{t\to\infty}\Omega_{t|t} = \Omega$. Consequently, $\lim_{t\to\infty}\mathsf{K}_t = \mathsf{K}$, and the stationary covariance matrix of the innovations is $\Sigma = \lim_{t\to\infty}\Sigma_{t|t} = x_1'\Omega x_1 + x_2'\Sigma_{v_1}x_2$. As we show next, the expressions for $\Omega, \mathsf{K}, \Sigma$ obtained in a constant coefficient model are the same as those asymptotically produced by a recursive least square estimator.

Example 6.6 *Consider estimating the constant (steady state) real interest rate α_t using T observations on the nominal interest rate y_t, demeaned by the average inflation rate, where $y_t = \alpha_t + v_{1t}$ and v_{1t} is a martingale difference process with variance $\sigma^2_{v_1}$. An unbiased minimum variance estimator is $\hat{\alpha}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$. If y_{T+1} becomes available $\hat{\alpha}_{T+1} = \frac{1}{T+1}\sum_{t=1}^{T+1} y_t = \frac{T}{T+1}(\frac{1}{T}\sum_{t=1}^{T} y_t) + \frac{1}{T+1}y_{T+1} = \frac{T}{T+1}\hat{\alpha}_T + \frac{1}{T+1}y_{T+1}$ which is a recursive least square estimator. This estimator weights previous and current observations using the number of available observations and does not forget: each observation gets equal weight regardless of the time elapsed since it was observed. A more informative way to rewrite this expression is $\hat{\alpha}_{T+1} = \hat{\alpha}_T + \frac{1}{T+1}(y_{T+1} - \hat{\alpha}_T)$ and $\epsilon_t \equiv (y_{T+1} - \hat{\alpha}_T)$ is the innovation in forecasting y_{T+1}. Clearly, $\mathsf{K}_{T+1} = \frac{1}{T+1} \to 0$ as $T \to \infty$. Hence, as $T \to \infty$, $\hat{\alpha}_{T+1} \to \hat{\alpha}_T$.*

The recursions in (6.3)-(6.8) assume constant coefficients. The Kalman filter, however, can also be applied to models with time varying coefficients, as long as they are linear in parameters. For example, in the multivariate model

$$\begin{aligned} y_t &= \alpha_t y_{t-1} + v_{1t} \\ \alpha_t &= \alpha_{t-1} + v_{2t} \end{aligned} \tag{6.13}$$

recursive estimates of $\alpha_{t|t}$ and of the forecast error $\epsilon_t = y_t - \alpha_{t|t-1}y_{t-1}$ consistent with the information available at each t can be easily obtained. We extensively use models like (6.13) in chapter 10 when studying time varying Bayesian VAR models.

Exercise 6.8 *Consider the model $y_t = x_t'\alpha_t + v_{1t}$ where $\alpha_t = (I - \mathbb{D}_1)\alpha_0 + \mathbb{D}_1\alpha_{t-1} + v_{2t+1}$, α_0 is a constant; v_{1t} is a martingale difference with variance $\sigma^2_{v_1}$ and v_{2t} is a vector of martingale difference with variance Σ_{v_2}. Define $\alpha_t^\dagger = \alpha_t - \alpha_0$. Show the form of the updating equations for $\alpha_t^\dagger$ and Ω_t, assuming $\alpha_1^\dagger \sim \mathbb{N}(\alpha_{1|0}, \Omega_{1|0})$.*

A modified version of the Kalman filter can also be used in special nonlinear state space models; for example, those displaying structures like the one of exercise 6.3. To compute the Kalman gain in this case it is necessary to linearize the extended state space around the current estimate. For example, the updating equations are

$$\begin{aligned} \alpha_{t|t} &= \alpha_{t|t-1}\theta_{t|t-1} + \mathsf{K}_{1t}(y_t - \alpha_{t|t-1}) \\ \theta_{t|t} &= \theta_{t|t-1} + \mathsf{K}_{2t}(y_t - \alpha_{t|t-1}) \end{aligned} \tag{6.14}$$

where where $\mathsf{K}_{1t}, \mathsf{K}_{2t}$ are matrices involving linear and quadratic terms in the predictors $\theta_{t|t-1}$ and $\alpha_{t|t-1}$, linear terms in the variance $\sigma^2_{v_1}$ and in past Kalman gains (see Ljung and Soderstroem (1983), pp. 39-40 for details).

If initial conditions and innovations are normally distributed, the Kalman filter predictor is the best in both the class of linear and nonlinear predictors. Moreover, forecasts of y_t are normal with mean $x_1'\alpha_{t|t-1}$ and variance $\Sigma_{t|t-1}$. When the two above conditions are not satisfied, the Kalman filter only produces the best linear predictor for y_t, based on

information at time t. That is, there are nonlinear filters which produce more efficient estimators than those produced in (6.5)-(6.6). A nonlinear filter for a model with binomial innovations was described in chapter 3 (see also Hamilton (1994), ch.22).

Example 6.7 *As we have seen, a two-state Markov switching model for y_t can be written as $y_t = a_0 + a_1\varkappa_t + y_{t-1}$ where $\varkappa_t$ has an AR(1) representation of the form*

$$\varkappa_t = (1 - p_2) + (p_1 + p_2 - 1)\varkappa_{t-1} + v_{1t} \tag{6.15}$$

and where v_{1t} can take four possible values $[1 - p_1, -p_1, -(1 - p_2), p_2]$ with probabilities $[p_1, 1-p_1, p_2, 1-p_2]$ and therefore is non-normal. It is immediate to verify that this process has a state space representation and that the orthogonality assumptions needed for identification are satisfied. However, while $corr(v_{1t}, \varkappa_{t-\tau}) = 0\ \forall\tau > 0$, the two processes are not independent. Equation (6.15) can be rewritten as

$$(1-(p_1+p_2-1)\ell)\Delta y_t = a_1(1-(p_1+p_2-1)\ell)\varkappa_t = a_1(1-p_2)+a_0(2-p_1-p_2)+v_{1t} \tag{6.16}$$

Hence, although y_t has a linear ARIMA(1,1,0) structure, Kalman filter estimates of $y_{t+1|t}$ based on such a model are suboptimal since the non-linear structure present in v_{1t} is ignored. In fact, optimal forecasts are obtained using

$$E_t\Delta y_{t+1} = a_0 + a_1 E_t\varkappa_{t+1} = a_0 + a_1[\frac{1-p_2}{2-p_1-p_2} + (p_1+p_2-1)(P[\varkappa_t = 1|\mathcal{F}_t] - \frac{1-p_2}{2-p_1-p_2})] \tag{6.17}$$

where $\mathcal{F}_t$ represents the information set at t. The nonlinear filtering algorithm described in chapter 3 uses (6.17) to obtain estimates of $\varkappa_t$.

While we have assumed that the measurement error and the error in the state equation are uncorrelated, in some situations this assumption may be unpalatable. For example, in the context of a model like (6.13), one may want to have the innovations in y_t and in α_t to be correlated. Relaxing this assumption requires some ingenuity. The next exercise shows that a system with a serially correlated measurement error is equivalent to a system with correlation between innovations in the transition and the measurement equations.

Exercise 6.9 *Suppose that all coefficients are constant, that $\mathbb{D}_0 = 0$ and that v_t in equation (6.1) satisfies $v_{1t} = \rho_v v_{1t-1} + \upsilon_t$ where ρ_v has all the eigenvalues less than one in absolute value and υ_t is a martingale difference with covariance matrix Σ_υ. Assuming that $E(v_{2t}\upsilon'_\tau) = 0\ \forall t$, and $\tau \neq t$, show that an equivalent state space representation is given by (6.2) and by $y_t^\dagger = x_{1t}^\dagger\alpha_t + v_{1t+1}^\dagger$ where $y_t^\dagger = y_{t+1} - \rho_v y_t$, $x_{1t}^\dagger = x_{1t}\mathbb{D}_1 - \rho_v x_{1t}$ and $v_{1t+1}^\dagger = x_{1t}\mathbb{D}_2 v_{2t+1} + v_{1t+1}$.*

Exercise 6.10 *Suppose α_t is normally distributed with mean $\bar{\alpha}$ and variance $\bar{\Sigma}_\alpha$, that $y_t = x_1'\alpha_t + v_{1t}$, where v_{1t} is orthogonal to α_t, and $v_{1t} \sim iid\ \mathbb{N}(0, \sigma_{v_1})$.*
(i) Show that $y_t \sim \mathbb{N}(x_1'\bar{\alpha}, x_1'\bar{\Sigma}_\alpha x_1 + \sigma_{v_1}^2)$.

(ii) Using the fact that the posterior density of α_t *is* $g(\alpha_t|y_t) = \frac{g(\alpha_t)f(y_t|\alpha_t)}{g(y_t)}$*, show that* $g(\alpha_t|y_t) \propto exp\{-0.5((\alpha_t - \bar{\alpha})'\bar{\Sigma}_\alpha^{-1}(\alpha_t - \bar{\alpha}) + (y_t - x_1'\alpha_t)'\sigma_{v_1}^{-2}(y_t - x_1'\alpha_t)\} \equiv exp\{-0.5((\alpha_t - \tilde{\alpha})'\tilde{\Sigma}_\alpha^{-1}(\alpha_t - \tilde{\alpha})\}$ *where* $\tilde{\alpha} = \bar{\alpha} + \bar{\Sigma}_\alpha x_1\sigma_{v_1}^{-2}(y_t - x_1'\bar{\alpha})$ *and* $\tilde{\Sigma}_\alpha = \bar{\Sigma}_\alpha + \bar{\Sigma}_\alpha x_1\sigma_{v_1}^{-2}x_1'\bar{\Sigma}_\alpha$*.*

Exercise 6.11 *A generalized version of a log-linearized RBC model can be written as* $\alpha_t = \mathbb{D}_{1t-1}\alpha_{t-1} + v_{2t},\ v_{2t} \sim (0, \Sigma_t)$*, and* $y_t = x_{1t}'\alpha_t$ *where* α_t *represents a vector of states and shocks and* y_t *are the controls. Assume that* $\Sigma_t, x_{1t}, \mathbb{D}_{1t-1}$ *are known.*
(i) Find the updating equation for the forecast error variance and show that $x_{1t}'\Omega_{t|t}x_{1t} = 0$*.*
(ii) Show that $\Omega_{t+1|t} = \mathbb{D}_{1t}\Omega_{t|t}\mathbb{D}_{1t}' + \Sigma_t$*.*

Given the recursive nature of Kalman filter estimates, it is easy to compute multistep forecasts of y_t. We leave the derivation of these forecasts as an exercise for the reader.

Exercise 6.12 *Consider the model (6.1)-(6.2) and the prediction of* $y_{t+\tau}$*. Show that the* τ*-steps ahead forecast error is* $x_{1t+\tau}'(\alpha_{t+\tau} - \alpha_{t+\tau,t}) + x_{2t+\tau}'v_{1t+\tau}$ *and that the MSE of the forecast is* $x_{1t+\tau}'\Omega_{t+\tau|t}x_{1t+\tau} + x_{2t+\tau}'\Sigma_{v_1}x_{2t+\tau}$*. Show the form of* $\alpha_{t+\tau|t}$ *and* $\Omega_{t+\tau|t}$*.*

Example 6.8 *Consider an* $m \times 1$ *VAR(q) model,* $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$*. As we have seen in example 6.1, this is a state space model for* $x_{1t}' = I,\ \alpha_t = y_t,\ \mathbb{D}_{1t} = \mathbb{A},\ \Sigma_{v_1} = 0,\ v_{2t} = E_t,\ \mathbb{D}_{2t} = I, \mathbb{D}_{0t} = 0$*. The* τ*-steps ahead forecast of* y_t *is* $E_t[y_{t+\tau}] = \mathbb{S}\mathbb{A}^\tau\mathbb{Y}_t$*, where* $\mathbb{S}$ *is a selection matrix. The forecast error variance is* $(\mathbb{S}(\mathbb{Y}_{t+\tau} - \mathbb{A}^\tau\mathbb{Y}_t))(\mathbb{S}(\mathbb{Y}_{t+\tau} - \mathbb{A}^\tau\mathbb{Y}_t)'))$*.*

6.2 The Prediction error decomposition of likelihood

Maximum likelihood estimation of nonlinear models is complicated. However, even in models like (6.1)-(6.2), which are conditionally linear in the parameters, maximization of the likelihood function is problematic when observations are not independent. This section is concerned with the practical question of constructing the likelihood function for models which have a format like (6.1)-(6.2), when y_t is serially correlated over time. It turns out that there is a convenient format, called prediction error decomposition, which can be used to estimate ARMA, structural VARs and, as we will see, DSGE models.

To understand what this decomposition entitles let $f(y_1, \dots y_T)$ be the joint density of $\{y_t\}_{t=1}^T$. Given the properties of joint densities, it is possible to decompose $f(y_1, \dots y_T)$ into the product of a conditional and a marginal, and repeatedly substituting we have:

$$\begin{aligned} f(y_1, \dots y_T) &= f(y_T|y_{T-1} \dots y_1)f(y_{T-1}, \dots y_1) \\ &= f(y_T|y_{T-1} \dots y_t)f(y_{T-1}|y_{T-2}, \dots y_1)f(y_{T-2}, \dots y_1) \\ &\dots \\ &= \prod_{j=0}^{J} f(y_{T-j}|y_{T-j-1} \dots y_1)f(y_1) \end{aligned} \tag{6.18}$$

and $\log f(y_1, \dots y_T) = \sum_j \log f(y_{T-j}, |y_{T-j-1} \dots y_1) + \log f(y_1)$. If $y = [y_1, \dots y_T] \sim \mathbb{N}(\bar{y}, \Sigma_y)$

$$\mathcal{L}(y|\phi) = \log f(y_1, \dots, y_T|\phi) = -\frac{T}{2}(\log 2\pi + \log|\Sigma_y|) - \frac{1}{2}(y-\bar{y})\Sigma_y^{-1}(y-\bar{y}) \tag{6.19}$$

where $\phi = (\bar{y}, \Sigma_y)$. Calculation of (6.19) requires the inversion of Σ_y, which is a $T \times T$ matrix, and this may be complicated when T is large. Using decomposition (6.18), we can partition $\mathcal{L}(y_1, \dots, y_t|\phi) = \mathcal{L}(y_1, \dots, y_{T-1}|\phi)\mathcal{L}(y_t|y_{T-1}, \dots, y_1, \phi)$. When$\{y_t\}_{t=1}^T$ is normal, both the conditional and the marginal blocks are normal.

Let $y_{t|t-1}$ be a predictor of y_t using information up to $t-1$. The prediction error is $\epsilon_t = y_t - y_{t|t-1} = y_t - E(y_t|y_{t-1}, \dots, y_1) + E(y_t|y_{t-1}, \dots, y_1) - y_{t|t-1}$ and its Mean Square Error (MSE) is $E(\epsilon_t - E(\epsilon_t))^2 = E(y_t - E(y_t|y_{t-1}, \dots y_1))^2 + E(E(y_t|y_{t-1}, \dots y_1) - y_{t|t-1})^2$. The best predictor of y_t, i.e. the one that makes the MSE of the prediction error as small as possible, is obtained when $E(y_t|y_{t-1}, \dots, y_1) = y_{t|t-1}$. Given this choice, the MSE of ϵ_t, denoted by $\sigma^2_{\epsilon_t}$, equals the variance of $(y_t|y_{t-1}, \dots, y_1)$.

The conditional density of y_t given information at time $t-1$ can then be written as:

$$\mathcal{L}(y_t|y_{t-1}, \dots, y_1, \sigma^2_{\epsilon_t}) = -\frac{1}{2}\log(2\pi) - \log(\sigma_{\epsilon_t}) - \frac{1}{2}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \tag{6.20}$$

Since (6.20) is valid for any $t > 1$ using (6.18) we have that

$$\begin{aligned}
\mathcal{L}(y_1, \dots y_T|\sigma^2_{\epsilon_1}, \dots, \sigma^2_{\epsilon_T}) &= \sum_{t=2}^{T} \mathcal{L}(y_t|y_{t-1}, \dots, y_1, \sigma^2_{\epsilon_2}, \dots, \sigma^2_{epsilon_T}) + \mathcal{L}(y_1|\sigma^2_{\epsilon_1}) \\
&= -\frac{T-1}{2}\log(2\pi) - \sum_{t=2}^{T}\log\sigma_{\epsilon_t} - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \\
&- \frac{1}{2}\log(2\pi) - \log\sigma_{\epsilon_1} - \frac{1}{2}\frac{(y_1 - \bar{y}_1)^2}{\sigma^2_{\epsilon_1}}
\end{aligned} \tag{6.21}$$

where $\bar{y}_1$ is the unconditional predictor of y_1. (6.21) is the decomposition we were looking for. Three important aspects need to be emphasized. First, (6.21) can be computed recursively, since it only involves one step ahead prediction errors and their optimal MSE. This should be contrasted with (6.19) where the entire vector of y_t's is used. Second, both the best predictor $y_{t|t-1}$ and the MSE of the forecast $\sigma^2_{\epsilon_t}$ vary with time. Therefore, we have transformed a time invariant problem into a problem involving quantities that vary over time. Third, if y_1 is a constant, prediction errors are constant and exactly equal to the innovations in y_t.

Example 6.9 *Consider a univariate AR(1) process $y_t = Ay_{t-1} + e_t$, $|A| < 1$, where e_t is normal martingale difference process with variance σ^2_e. Let $\phi = (A, \sigma^2_e)$. Assume that the process has started far in the past but it has been observed only from $t = 1$ on. For any t, $y_{t|t-1} \sim \mathbb{N}(Ay_{t-1}, \sigma^2_e)$. Hence, the prediction error $\epsilon_t = y_t - y_{t|t-1} = y_t - Ay_{t-1} = e_t$.*

Moreover, since the variance of e_t is constant, also the variance of the prediction error is constant (from time $t = 2$ on). Setting $\bar{y} = 0$, $y_1 \sim \mathbb{N}(0, \frac{\sigma_e^2}{1-A^2})$ and

$$\begin{aligned}\mathcal{L}(\phi) &= \sum_{t=2}^{T} \mathcal{L}(y_t|y_{t-1}, \ldots, y_1, \phi) + \mathcal{L}(y_1|\phi) \\ &= -\frac{T}{2}\log(2\pi) - T\log(\sigma_e) - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - Ay_{t-1})^2}{\sigma_e^2} + \frac{1}{2}(\log(1-A^2) - \frac{(1-A^2)y_1^2}{\sigma_e^2})\end{aligned}$$

Hence $\sigma_{\epsilon_t}^2 = \sigma_e^2$ for all $t \geq 2$, while $\sigma_{\epsilon_1}^2 = \frac{\sigma_e^2}{1-A^2}$.

Exercise 6.13 *Consider the univariate model $y_{1t} = A_1(\ell)y_{1t-1} + D(\ell)e_t + A_2 y_{2t}$, where y_{2t} are exogenous variables, $A_1(\ell)$ is a polynomial of order q_1, $D(\ell)$ is a polynomial of order q_2. Find $y_{1t|t-1}$ and $\sigma_{\epsilon_t}^2$ in this case. Show the form of the log likelihood function assuming that the first $q = max[q_1, q_2 + 1]$ values of $y_t = [y_{1t}, y_{2t}]$ are constants.*

Taking the initial observations as given is convenient since it eliminates a source of nonlinearities. In general, nonlinearities do not allow to compute an analytical solution to the first order conditions of the maximization problem and the maximum of the likelihood must be located using numerical techniques. Conditioning on the initial observations makes the maximization problem trivial in many cases. Note also that, as $T \to \infty$, the contribution of the first observation to the likelihood becomes negligible. Therefore, exact and conditional maximum likelihood coincide if the sample is large. Furthermore, when the model has constant coefficients, the errors are normally distributed and the initial observations fixed, maximum likelihood and OLS estimators are identical (see chapter 4 in the case of a VAR). This would not be the case when a model features moving average terms (see example 6.11), since nonlinearities do not wash out, even conditioning on the initial observations.

Example 6.10 *Consider finding the ML estimator of the AR process described in example 6.9. Conditioning on y_1 the log likelihood of $(y_2, \ldots, y_T)$ is proportional to $\sum_{t=2}^{T}\{-\log(\sigma_e) - \frac{1}{2\sigma_e^2}(y_t - Ay_{t-1})^2\}$. Maximizing this quantity with respect to A (conditional on σ_e^2), is equivalent to minimizing $(y_t - Ay_{t-1})^2$, which produces $A_{ML} = A_{ols}$. Using A_{ML}, the likelihood can be concentrated to obtain $-\frac{T-1}{2}\log(\sigma_e^2) - \frac{\sum_t \epsilon_t'\epsilon_t}{2\sigma_e^2}$. Maximizing it with respect to σ_e^2 leads to $\sigma_{ML}^2 = \frac{\sum_t \epsilon_t'\epsilon_t}{T-1}$. Suppose now that we do not wish to condition on y_1. Then the likelihood function is proportional to $\sum_{t=2}^{T}\{-\log(\sigma_e) - \frac{1}{2\sigma_e^2}(y_t - Ay_{t-1})^2\} + \{-0.5\log(\frac{\sigma_e^2}{1-A^2}) - \frac{y_1^2(1-A^2)}{2\sigma_e^2}\}$. If $T \to \infty$, the first observation makes a negligible contribution to the likelihood of the sample. Therefore, conditional ML estimates of A asymptotically coincide with full ML estimates, provided $|A| < 1$.*

Consider, finally, the case where A is time varying, e.g. $A_t = \mathbb{D}_1 A_{t-1} + v_{2t}$. Conditional on some A_0, the recursive conditional maximum likelihood estimator of $A_{t|t}$ and the smoothed maximum likelihood estimator $A_{t|T}$ can be obtained with the Kalman filter and the Kalman smoother. As $T \to \infty$, the importance of the initial observation will be discounted as long as the roots of $\mathbb{D}_1$ are all less than one in absolute value.

Exercise 6.14 *(i) Suppose that $y_t = x_t'\alpha + e_t$ where e_t is normal martingale difference with variance $\sigma^2_{e_t}$ and let x_t be fixed regressors. Show how to derive the prediction error decomposition of the likelihood for this model.*
(ii) Let x_t be a random variable, normally distributed with mean $\bar{x}$ and variance Σ_x. Show how to compute the prediction error decomposition of the likelihood in this case.

Multivariate prediction error decompositions present no difficulties. If y_t is $m \times 1$ vector

$$\mathcal{L}(y|\phi) = -\frac{Tm}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log|\Sigma_{t|t-1}| - \frac{1}{2}\sum_{t=1}^{T}(y_t - y_{t|t-1})\Sigma^{-1}_{t|t-1}(y_t - y_{t|t-1}) \quad (6.22)$$

where $\epsilon_t = y_t - y_{t|t-1} \sim \mathbb{N}(0, \Sigma_{t|t-1})$ and where we assume $y_1 \sim \mathbb{N}(\bar{y}_1, \Sigma_{1|0})$ and $\epsilon_1 = y_1 - \bar{y}_1$.

Exercise 6.15 *Consider the setup of exercise 6.11. Show the form of $y_{t|t-1}$ and $\Sigma_{t|t-1}$ and the prediction error decomposition of the likelihood in this case.*

The prediction error decomposition is convenient in two respects. First, the building blocks of the decomposition are the forecast errors ϵ_t and their MSE $\Sigma_{t|t-1}$. Since the Kalman filter produces these quantities recursively, it can be used to build the prediction error decomposition of the likelihood of any model which has a state space format. Second, since any ARMA process has a state space format, the prediction error decomposition of the likelihood can be easily obtained for a variety of statistical an economic models.

Maximization of the likelihood conditional on the initial observations, can be obtained by extending algorithm 6.1. Let $\phi = [vec(x_1'), vec(x_2'), vec(\mathbb{D}_1), vec(\mathbb{D}_0), vec(\mathbb{D}_2), \Sigma_{v_1}, \Sigma_{v_2}]$. Then

Algorithm 6.2

1) *Choose some initial $\phi = \phi_0$.*

2) *Do steps 1)-4) of algorithm 6.1.*

3) *At each step save $\epsilon_t = y_t - y_{t|t-1}$ and $\Sigma_{t|t-1}$. Construct the log likelihood (6.22).*

4) *Update initial estimates of ϕ using any of the methods described in section 6.3.*

5) *Repeat steps 2)-4) until $|\phi^l - \phi^{l-1}| \leq \iota$; or $\frac{\partial \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi^l} < \iota$, or both, for ι small.*

Two comments on algorithm 6.2 are in order. First, the initial values of iterations can be typically obtained by running an OLS regression on the constant coefficient version of the model. If the assumptions underlying the state space specification are correct this will consistently estimate the average value of the parameters. Second, for large dimensional problems, maximization routines typically work better if Choleski factor of $\Sigma_{t|t-1}$, is used in the computations of the likelihood.

The conditional prediction error decomposition is particularly useful to estimate models with MA terms. Such models are difficult to deal with in standard setups but fairly easy to estimate within a state space framework.

Example 6.11 *In testing the efficiency of foreign exchange markets one runs a monthly regression of the realized three month change in spot exchange rate at $t+3$ on the forward premium quoted at t for $t+3$. As seen in chapter 5, such a regression has moving average errors of order up to 2 because of overlapping time intervals. Therefore, a model for testing efficiency could be $y_{t+3} = b_0 x_t + \epsilon_{t+3}$ with $\epsilon_{t+3} = e_{t+3} + b_1 e_{t+2} + b_2 e_{t+1}$ where, under the null hypothesis, $b_0 = 1$ and e_t is a normal martingale difference with variance σ_e^2. This model can be cast into a state space framework by defining $\mathbb{D}_0 = 0, \mathbb{D}_2 = I, x'_{2t} = I, v_{1t} = 0$,* $\alpha_t = \begin{bmatrix} x_t \\ e_{t+3} \\ e_{t+2} \\ e_{t+1} \end{bmatrix}$, $\mathbb{D}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$, $x_{1t} = \begin{bmatrix} b_0 \\ 1 \\ b_1 \\ b_2 \end{bmatrix}$ $v_{2t} = \begin{bmatrix} x_t \\ e_{t+3} \\ 0 \\ 0 \end{bmatrix}$. *Suppose we are interested in estimating $[b_0, b_1, b_2]$ and in testing $b_0 = 1$. Then ML estimates can be obtained starting the Kalman filter at $\alpha_{1|0} = [x_1, 0, 0, 0]$ and $\Omega_{1|0} = diag\{\sigma_x^2, \sigma_e^2, \sigma_e^2, \sigma_e^2\}$ where σ_x^2 is the unconditional variance of the forward premium and σ_e^2 could be either the variance of the error $\hat{e}_t = y_t - \hat{b}_0 x_{t-3}$ in a training sample (say, from $-\tau$ to 0) or set to an arbitrarily large number. To start the iterations we need x_{10}, that is, we need some initial estimates of (b_0, b_1, b_2). An estimate of b_0 could be obtained in a training sample or, if no such a sample exists, using available data but disregarding serial correlation in the error term. Initial estimates of b_1 and b_2 could then be $b_1 = b_2 = 0$. Then the sequence of iterations producing $\alpha_{t|t-1}$ and $\Omega_{t|t-1}$ can be used to compute the likelihood function. Note that for this simple problem one could evaluate the likelihood numerically at successive grids of, say, 20 points in each dimension and locate the maximum numerically.*

Exercise 6.16 *Consider an AR(2) process $y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + e_t$ where $e_t \sim iid\ \mathbb{N}(0, \sigma_e^2)$. Show that the exact log likelihood function is $\mathcal{L}(\phi) \propto -T\log(\sigma_e) + 0.5log((1 + A_2)^2[(1 - A_2)^2 - A_1^2]) - \frac{1+A_2}{2\sigma^2}[(1 - A_2)(y_1 - \bar{y})^2 - 2A_1(y_1 - \bar{y})(y_2 - \bar{y}) + (1 - A_2)(y_2 - \bar{y})^2] - \sum_{t=3}^{T} \frac{(y_t - A_0 - A_1 y_{t-1} + A_2 y_{t-2})^2}{2\sigma^2}$ where $\bar{y} = \frac{A_0}{1-A_1-A_2}$. Which terms disappear if a conditional likelihood approach is used? Show that $\sigma_{ML}^2 = \frac{1}{T-2}\sum_{t=3}^{T}(y_t - A_{0,ML} - A_{1,ML} y_{t-1} - A_{2,ML} y_{t-2})^2$.*

6.2.1 Some Asymptotics of ML estimators

It is fairly standard to show that, under regularity conditions, ML estimates of the parameters of a state space model are consistent and asymptotically normal (see e.g. Harvey (1991)). The conditions needed are generally of three types. First, we need the state equation to define a covariance stationary process. One simple sufficient condition for this is that the eigenvalues of $\mathbb{D}_{1t}$ are all less than one in absolute value for all t. Second, if the model includes exogenous variables we also need them to be covariance stationary, linearly regular processes. Third, we need the true parameters not to lie on the boundary of the parameter space. Then, under the above conditions, $\sqrt{T}(\phi_{ML} - \phi_0) \xrightarrow{D} \mathbb{N}(0, \Sigma_\phi)$ where $\Sigma_\phi = -T^{-1}E(\sum_t \frac{\partial^2 \mathcal{L}}{\partial\phi\partial\phi'}|_{\phi=\phi_0})^{-1}$.

For the case in which the innovations are the errors in the measurement equation, the asymptotic covariance matrix is block diagonal, as it is shown next.

Example 6.12 *For an AR(1) model it is quite easy to derive Σ_ϕ. In fact, conditional on the initial observations, the log likelihood is $\mathcal{L}(\phi) \propto -\frac{T-1}{2}\log\sigma_\epsilon^2 - \frac{\sum_{t=2}^{T}\epsilon_t^2}{2\sigma_\epsilon^2}$ where $\epsilon_t = y_t - Ay_{t-1}$ and the matrix of second derivatives is $\begin{bmatrix} -\sigma_\epsilon^{-2}\sum_t y_{t-1}^2 & -\sigma_\epsilon^{-4}\sum_t \epsilon_t y_{t-1} \\ -\sigma_\epsilon^{-4}\sum_t \epsilon_t y_{t-1} & (2\sigma_\epsilon^4)^{-1}(T-1) - \sigma_\epsilon^{-6}\sum_t \epsilon_t^2 \end{bmatrix}$. Since the expectation of the off-diagonal elements is zero, the asymptotic covariance matrix is diagonal with $var(A) = \frac{\sigma_\epsilon^2}{(T-1)\sum_t y_{t-1}^2}$ and $var(\sigma_e^2) = \frac{2\sigma_\epsilon^4}{T-1}$.*

The derivation of the Kalman filter assumes that the innovations in the measurement and in the observation equations are normally distributed. Since the likelihood function is calculated with the Kalman filter estimates, one may wonder what are the properties of maximum likelihood estimates when the distribution of the driving forces is misspecified.

As mentioned, misspecification of the distribution of the errors does not create consistency problems for Kalman filter estimates. It turns out that this property carries over to maximum likelihood estimates. In fact, maximum likelihood estimates obtained incorrectly assuming a normal distribution (typically called quasi-ML) have nice properties under a set of regularity conditions. We ask the reader to verify that this is the case for a simple problem in the next exercise.

Exercise 6.17 *Suppose observations on y_t are drawn from a t-distribution with a small number of degrees of freedom (say, less than 5) but that an econometrician estimates the constant coefficient state space model $y_t = \alpha_t + v_{1t}$, $\alpha_t = \alpha_{t-1}$ where v_{1t} is a normal martingale difference with variance $\sigma_{v_1}^2$. Show that the ML estimator for α_t based on the wrong (normal) distribution will be consistent and asymptotically normal. Show the form of the asymptotic covariance matrix.*

Intuitively, if the sample size is large and homogeneous, a normal approximation is appropriate. In the context of a constant coefficient state space model, we could have achieved the same conclusion by noting that recursive OLS is consistent and asymptotically normal if the regressors are stationary, ergodic and uncorrelated with the errors and that recursive OLS and Kalman filter-ML estimates coincide if a conditional likelihood is used.

When the coefficients of the state space model are time varying, ML estimates obtained with misspecified errors are no longer asymptotically equivalent to those of the correct model and Kalman filter estimates are not best linear MSE estimates of α_t.

We have seen that maximum likelihood estimates have an asymptotic covariance matrix equal to which is $-\frac{1}{T}E(\frac{\partial^2\mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_0})^{-1}$. There are many ways to estimate this matrix. One is to evaluate the quantity at the ML estimator, substituting averages for expectations, that is, $\text{var}_1(\phi) = (-\sum_t \frac{\partial^2\mathcal{L}_t(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_{ML}})^{-1}$. An alternative is obtained noting that an approximation to the second derivatives of the likelihood function can be calculated taking the derivatives of the scores, i.e. $\text{var}_2(\phi) = (\sum_t(\frac{\partial\mathcal{L}_t(\phi)}{\partial\phi}|_{\phi=\phi_{ML}})(\frac{\partial\mathcal{L}_t(\phi)}{\partial\phi}|_{\phi=\phi_{ML}})')^{-1}$.

Finally, a quasi ML estimator can be obtained combining the two above estimators. That is, $\text{var}_3(\phi) = -((var_1(\phi))(var_2(\phi))^{-1}(var_1(\phi))$.

Exercise 6.18 *For the AR(1) model considered in example 6.12, show the form of the three estimates of the asymptotic covariance matrix.*

Hypothesis testing on the parameters is fairly standard. Given the asymptotic normality of ML estimates, one could use t-tests to verify simple restrictions on the parameters or likelihood ratio tests when more general hypotheses are involved.

Example 6.13 *Continuing with example 6.11, to test $b_0 = 1$ use $\frac{b_{0,ML}-1}{\sigma_{b_{0,ML}}}$ and compare it to a t distribution with $T-1$ degrees of freedoms (or to a normal (0,1), if T is large). Alternatively, one could estimate the model under the restriction $b_0 = 1$, construct the likelihood function, calculate $-2[\mathcal{L}(b_{0,ML}) - \mathcal{L}(b_0 = 1)]$ and compare it with a $\chi^2(1)$.*

As we have seen with GMM, it may be more convenient at times to use estimates of a restricted model. This would be the case, for example, if the model is non-linear, but it becomes linear under some restrictions, or if it contains MA terms. In this case one can use the Lagrangian Multiplier (LM) statistic $\frac{1}{T}[\sum_t(\frac{\partial\mathcal{L}(\phi)}{\partial\phi})|_{\phi^{re}}]'\Sigma_\phi^{-1}[(\sum_t \frac{\partial\mathcal{L}(\phi)}{\partial\phi})_{\phi^{re}})] \sim \chi^2(\nu)$, where ν is the number of restrictions.

Example 6.14 *For the model of example 6.2, if $D_1 = 0$, conditional ML estimates of $A = [A_1, A_2]'$ solve the normal equations $Ax'x = x'y$ where $x_t = [y_{t-1}, y_{t-2}]$, $x = [x_1, \dots x_t]'$. However, if $D_1 \neq 0$ the normal equations are nonlinear and no analytical solution exists. Therefore, one may impose $D_1 = 0$ for estimation and test if the restriction holds.*

Two non-nested hypotheses can be evaluated using, for example, forecasting accuracy tests like the one of Diebold and Mariano (1995). Let ϵ_t^i be the prediction errors produced by specification $i = 1, 2$ and let $h_t = (\epsilon_t^1)^2 - (\epsilon_t^2)^2$. Then, under the hypothesis of similar predictive accuracy, the statistic $\mathsf{S} = \frac{\bar{h}}{se(h)}$, where $\bar{h} = \frac{1}{T}\sum_t h_t$, $se(h) = \sqrt{\frac{1}{T}\sum_t(h_t - \bar{h})^2}$ is asymptotically normally distributed with mean zero and variance one. We will use this statistic in section 6.5 when comparing the forecasting accuracy of a DSGE model relative to an unrestricted VAR.

6.3 Numerical tips

There are many ways to update initial estimates in step 4) of algorithm 6.2. Here we only briefly list some of them and highlighting advantages and disadvantages of each.

- Grid search.
  This maximization method is feasible when the dimension of ϕ is small. It involves discretizing the problem and selecting the value of ϕ which achieves the maximum on the grid. One advantage of the approach is that no derivatives of the likelihood are

needed - which can be useful if the problem is complicated. When the likelihood is globally concave, the approach will find an approximation to the maximum. However, if multiple peaks are present, it may select local maxima. For this reason, the grid should be fine enough to avoid pathologies. While care should be exercised in taking them as final estimates, grid estimates are useful as initial conditions for other algorithms.

- Simplex method
  A k-dimensional simplex is spanned by $k+1$ vectors which are the vertices of the simplex (e.g. if $k = 2$, two dimensional simplexes are triangles). This method is typically fast and works as follows. If a maximum is found at some iteration, the method substitutes it with a point on the ray from the maximum through the centroid of the remaining points. For example, if $\mathcal{L}(\phi_m) = \max_{j=1,k+1} \mathcal{L}(\phi_j)$, we replace ϕ_m by $\varrho\phi_m + (1-\varrho)\bar{\phi}$, where $\bar{\phi}$ is the centroid, $0 < \varrho < 1$ and repeat the maximization. This approach does not require the calculation of gradients or second derivatives of the likelihood and can be used when other routines fail. The major disadvantage is that no standard errors for the estimates are available.

- Gradient methods
  All algorithms in this class update initial estimates by taking a step based on the gradient of the likelihood at the initial estimate. They differ in the size and the direction in which the step is taken.

  a) Method of Steepest ascent.
  At each iteration l, parameters are updated using: $\phi^l = \phi^{l-1} + \frac{1}{2\lambda} gr(\phi^l)$ where $gr(\phi^l) = \frac{\partial \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi^l}$ and λ is the Lagrangian multiplier of the problem $\max_{\phi^l} \mathcal{L}(\phi^l)$ subject to $(\phi^l - \phi^{l-1})'(\phi^l - \phi^{l-1}) = \kappa$. In words, the method updates current estimates using the scaled gradient of the likelihood. λ is a smoothness parameter which prevents large jumps in ϕ between iterations (it plays the same role as λ in the Hodrick Prescott or exponential smoothing filters). Note that if $\phi^l \approx \phi^{l-1}$, $gr(\phi^l) \approx gr(\phi^{l-1})$ and one can use $\phi^l = \phi^{l-1} + \varrho gr(\phi^{l-1})$ where ϱ is small positive scalar (e.g. 10^{-5}). This choice is very conservative and avoids jumps in the estimates. However, a lot of iterations are typically needed before convergence is achieved and convergence could only be to local maximum. It is therefore a good idea to start the algorithm from several initial conditions and check whether the same maximum is obtained.

  b) Newton-Raphson Method
  The method is applicable if $\frac{\partial^2 \mathcal{L}(\phi)}{\partial \phi \partial \phi'}$ exists and if $\mathcal{L}(\phi)$ is concave (i.e. the matrix of second derivatives is positive definite). In this case, taking a second order expansion of $\mathcal{L}(\phi)$ around ϕ_0, we have:

$$\mathcal{L}(\phi) = \mathcal{L}(\phi_0) + gr(\phi_0)(\phi - \phi_0) - 0.5(\phi - \phi_0)' \frac{\partial^2 \mathcal{L}(\phi)}{\partial \phi \partial \phi'} (\phi - \phi_0) \tag{6.23}$$

Maximizing (6.23) with respect to ϕ and using ϕ^{l-1} as an estimate of ϕ_0 we have

$$\phi^l = \phi^{l-1} + (\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi^{l-1}})^{-1} gr(\phi^{l-1}) \tag{6.24}$$

If likelihood is quadratic, (6.24) generates convergence in one step. If it is close to quadratic, iterations on (6.24) will converge quickly and the global maximum will be achieved. However, if the likelihood is far from quadratic, not globally concave or if ϕ_0 is far away from the maximum, the method may have worse properties than the method of steepest ascent. Note that $(\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'})^{-1}$ can be used to provide an estimate of the variance covariance matrix of ϕ at each iteration. Once could combine steepest-ascent and Newton-Raphson methods into an hybrid one which shares the good properties of both, may speed up calculation without producing large jumps in the parameters estimates. This is done, e.g., by choosing $\phi^l = \phi^{l-1} + \varrho(\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi^{l-1}})^{-1} gr(\phi^{l-1})$ where $\varrho > 0$ is a small scalar.

c) Modified Newton-Raphson.

The basic Newton-Raphson method requires the calculation of the matrix $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ and its inversion. When ϕ is of large dimension this maybe computationally difficult. The modified Newton-Raphson method uses the fact that $\frac{\partial gr(\phi)}{\partial\phi} \approx \frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ and guesses the shape of $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ at the existing estimate using the derivative of the gradient. Let Σ^l be an estimate of $[\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}]^{-1}$ at iteration l. Then the method updates estimates of ϕ using (6.24) where

$$(\Sigma^l) = (\Sigma^{l-1}) + \frac{(-\rho^l \Sigma^{l-1} gr^{l-1})(-\rho^l \Sigma^{l-1} gr^{l-1})'}{(-\rho^l \Sigma^{l-1} gr^{l-1})\Delta gr^l} - \frac{\Sigma^{l-1}\Delta gr^l (\Delta gr^l)'(\Sigma^{l-1})^{-1}}{(\Delta gr^l)'\Sigma^{l-1}\Delta gr^l}$$

and $\Delta\phi^l = \phi^l - \phi^{l-1}$, $\Delta gr(\phi^l) = gr(\phi^l) - gr(\phi^{l-1})$. If likelihood is quadratic and the number of iterations large, $\lim_{l\to\infty} \phi^l = \phi_{ML}$ and $\lim_{l\to\infty} \Sigma^l = (\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}|_{\phi=\phi_{ML}})^{-1}$. Standard errors of the estimate can be read off the diagonal elements of Σ^l evaluated at ϕ_{ML}.

d) Scoring Method.

This method uses the information matrix $E\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ in place of $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ in the calculation where the expectation is evaluated at $\phi = \phi^{l-1}$. The information matrix approximation is convenient since it typically has a simpler expression than the Hessian.

e) Gauss-Newton-scoring method.

The Gauss-Newton method uses a function of $(\frac{\partial e}{\partial\phi}|_{\phi=\phi^l})'(\frac{\partial e}{\partial\phi}|_{\phi=\phi^l})$ as an approximation to $\frac{\partial^2 \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$, where ϕ_0^l is the value of ϕ at iteration l and e_t is the vector of errors in the model. In the case of constant state space models, the approximation is proportional to the vector of regressors constructed using the right hand side variables of both the state and the measurement equations. When the model is linear, Gauss-Newton and scoring approximations are identical.

6.4 ML estimation of DSGE models

Maximum likelihood estimation of the parameters of a DSGE model is a straightforward application of the methods we have described so far. As we have already seen in chapter 2, the log linearized solution of a DSGE model is of the form:

$$y_{2t} = \mathcal{A}_{22}(\theta)y_{2t-1} + \mathcal{A}_{23}(\theta)y_{3t} \quad (6.25)$$

$$y_{1t} = \mathcal{A}_{12}(\theta)y_{2t-1} + \mathcal{A}_{13}(\theta)y_{3t} \quad (6.26)$$

where y_{2t} includes the states and the driving forces, y_{1t} all other endogenous variables and y_{3t} the shocks of the model. Here $\mathcal{A}_{ii'}(\theta), i, i' = 1, 2$ are time invariant (reduced form) matrices which depend on $\theta = (\theta_1, \ldots, \theta_k)$, the structural parameters of preferences, technologies and government policies. Note also that there are cross equation restrictions in the sense that some $\theta_j, j = 1, \ldots, k$ may appear in more than one entry of these matrices.

Example 6.15 *In the working capital model considered in exercise 1.14 of chapter 2, setting $K_t = 1, \forall t$, y_{2t} includes lagged real balances $\frac{M_{t-1}}{p_{t-1}}$ and lagged deposits dep_{t-1}; y_{3t} includes shocks to the technology ζ_t and to the monetary rule M_t^g, while y_{1t} includes all the remaining endogenous variables (hours (n_t), output (GDP_t), the nominal interest rate (i_t) and the inflation rate (π_t)). Setting $N^{ss} = 0.33$, $\eta = 0.65$, $\pi^{ss} = 1.005$, $\beta = 0.99$, $(\frac{c}{GDP})^{ss} = 0.8$, the persistence of the shocks to 0.95 and the parameters of the policy rule to $a_2 = -1.0; a_1 = 0.5; a_3 = 0.1, a_0 = 0$, the log-linearizing solution has the following state space representation:*

$$\begin{bmatrix} \widehat{\frac{M_t}{p_t}} \\ \widehat{dep_t} \\ \widehat{GDP_t} \\ \widehat{n}_t \\ \widehat{i}_t \\ \widehat{\Pi}_t \end{bmatrix} = \begin{bmatrix} -0.4960 & 0.3990 \\ -1.0039 & 0.8075 \\ -0.3968 & 0.3192 \\ 0.9713 & -0.7813 \\ 2.0219 & -1.6264 \end{bmatrix} \begin{bmatrix} \widehat{\frac{M_{t-1}}{p_{t-1}}} \\ \widehat{dep_{t-1}} \end{bmatrix} + \begin{bmatrix} 1.3034 & -0.1941 \\ 1.1459 & -1.4786 \\ 1.0427 & -0.1552 \\ -0.3545 & 0.3800 \\ -0.9175 & -1.2089 \end{bmatrix} \begin{bmatrix} \widehat{\zeta_t} \\ \widehat{M^g}_t \end{bmatrix}$$

While in example 6.15 we have chosen a log-linear approximation, DSGE models with quadratic preferences and linear constraints also fit into this structure (see e.g. Hansen and Sargent (1998)). In fact, (6.25)-(6.26) are very general, do not require any certainty equivalence principle to obtain and need not be the solution to the model, as the next example shows.

Example 6.16 *(Watson) Suppose a model delivers the condition $E_t y_{t+1} = \alpha y_t + x_t$ where $x_t = \rho x_{t-1} + e_t^x$, x_0 given. This could be, e.g., a New-Keynesian Phillips curve, in which case x_t are marginal costs, or stock price relationship, in which case x_t are dividends. Using the innovation representation we have $x_t = E_{t-1}x_t + e_t^x$, $y_t = E_{t-1}y_t + e_t^y$ where $E_t x_{t+1} = \rho x_t = \rho(E_{t-1}x_t + e_t^x)$ and $E_t y_{t+1} = \alpha y_t + x_t = \alpha(E_{t-1}y_t + e_t^y) + (E_{t-1}x_t + e_t^x)$. Letting $y_{1t} = [x_t, y_t]$, $y_{2t} = [E_t x_{t+1}, E_t y_{t+1}]$, $y_{3t} = [e_t^x, v_t]$, where $v_t = e_t^y - E(e_t^y | e_t^x) = e_t^y - \kappa e_t^x$, $\mathcal{A}_{11}(\theta) = I$, $\mathcal{A}_{12}(\theta) = \begin{bmatrix} 1 & 0 \\ \kappa & 1 \end{bmatrix}, \mathcal{A}_{22}(\theta) = \begin{bmatrix} \rho & 0 \\ 1 & \alpha \end{bmatrix}, \mathcal{A}_{21}(\theta) = \begin{bmatrix} \rho & 0 \\ 1+\alpha\kappa & \alpha \end{bmatrix}$,*

it is immediate to see that the model fits into (6.25)-(6.26). Here the parameters to be estimated are $\theta = (\alpha, \rho, \kappa, \sigma_e^2, \sigma_v^2)$.

In general, one has two alternatives to derive a representation which fits (6.25)-(6.26): solve the model, as we have done in example 6.15, or use the rational expectations assumption, as we have done in example 6.16,

Exercise 6.19 *Consider a version of a consumption-saving problem where consumers are endowed with utility of the form* $u(c) = \frac{c^{1-\varphi}}{1-\varphi}$, *the economy is small relative to the rest of world and the resource constraint is* $c_t + B_t \leq GDP_t + (1 + r_t)B_{t-1}$ *where* B_t *are internationally traded bonds and* r_t *is the net real interest rate, taken as given by the agents.*
(i) Derive a log linearized version of the Euler equation and show how to map it into the framework described in example 6.16.
(ii) Show the entries of the matrices in the state space representation.
(iii) How would you include a borrowing constraint $B_t < \bar{B}$ *in the setup?*

Exercise 6.20 *Consider the labor hoarding model studied in exercise 4.1 of chapter 5 where agents have preferences over consumption, leisure and effort and firms distinguish between labor and effort in the production function. Cast the log-linearized Euler conditions into a state space framework using an innovation representation.*

Clearly, (6.25)-(6.26) are in a format estimable with the Kalman filter. In fact, recursive estimates of y_{2t} can be obtained, given some initial conditions y_{20}, if $\mathcal{A}_{ii'}(\theta)$, $\sigma_{y_3}^2$ are known. Given these recursive estimates forecast errors can be computed. Hence, for each choice of θ, we can calculate the likelihood function via the prediction error decomposition and update estimates using one of the algorithms described in section 3. Standard errors for the estimated parameters can be read off the Hessian, evaluated at maximum likelihood estimates, or any approximation to it.

Despite the simplicity of this procedure, there are several issues, specific to DSGE models, one must deal with when using ML to estimate structural parameters. The first has to do with the number of series used in the estimation. As it is clear from (6.25)-(6.26), the covariance matrix of the vector $[y_{1t}, y_{2t}]$ is singular, a restriction unlikely to hold in the data. This singularity was also present in the innovation representation (6.11). Two options are available to the applied investigator: she can either select as many variables as there are shocks or artificially augment the space of shocks with measurement errors. For example, if the model is driven by a technology and a government expenditure shock, one selects two of (the many) series belonging to $[y_{1t}, y_{2t}]$ to estimate parameters. Kim (2000) and Ireland (2000) use such an approach in estimating versions of sticky price models. While this leaves some arbitrariness in the procedure, some variables may have little information about the parameters. Although a-priori it may hard to know which equations carry information, one could try to select variables so has to maximize the identifiability of the parameters. Alternatively, since some variables may not satisfy the assumptions needed to obtain consistent estimates (for example, they display structural breaks), one could choose the variables that are more likely to satisfy these assumptions.

Example 6.17 *In a standard RBC model driven by technology disturbances we have that* $[\hat{N}_t, \widehat{gdp}_t, \hat{c}_t]$ *are statically related to the states* $[\hat{K}_t, \hat{\zeta}_t]$ *via a matrix* $\mathcal{A}_{12}(\theta)$ *where* $\hat{\cdot}$ *refers to deviations from the steady state. Since the number of shocks is less than the number of endogenous variables, there are linear combinations of the controls which are perfectly predictable. For example, using equation (6.25) into (6.26) we have that* $\alpha_1 \hat{N}_t + \alpha_2 \widehat{gdp}_t + \alpha_3 \hat{c}_t = 0$ *where* $\alpha_1 = \mathcal{A}_1^{11}\mathcal{A}_1^{32} - \mathcal{A}_1^{12}\mathcal{A}_1^{31}$, $\alpha_2 = \mathcal{A}_1^{22}\mathcal{A}_1^{31} - \mathcal{A}_1^{21}\mathcal{A}_1^{32}$, $\alpha_3 = \mathcal{A}_1^{22}\mathcal{A}_1^{11} - \mathcal{A}_1^{31}\mathcal{A}_1^{21}$. *Similarly, using the equations for* $\widehat{gdp}_t, \hat{c}_t$ *and the law of motion of the capital stock we have* $\alpha_4 \hat{c}_t + \alpha_5 \hat{c}_{t-1} - \alpha_6 \widehat{gdp}_t - \alpha_7 \widehat{gdp}_{t-1} = 0$ *where* $\alpha_4 = \mathcal{A}_1^{12} + \delta[1 - \delta(\frac{K}{N})^{\eta}](\mathcal{A}_1^{12}\mathcal{A}_1^{31} - \mathcal{A}_1^{11}\mathcal{A}_1^{32}/[1 - \delta(\frac{K}{N})^{\eta}]$, $\alpha_5 = (1-\delta)\mathcal{A}_1^{12}$ $\alpha_6 = \mathcal{A}_1^{32} - \delta(\mathcal{A}_1^{12}\mathcal{A}_1^{31} - \mathcal{A}_1^{12}\mathcal{A}_1^{32})/[1 - \delta(\frac{K}{N})^{\eta}]$, $\alpha_7 = (1-\delta)\mathcal{A}_1^{32}$. *This implies that the system is stochastically singular and for any sample size the covariance matrix of the data is postulated to be of reduced rank.*

Attaching measurement errors to (6.26) is the option taken by Sargent (1979), Altug (1989) or McGrattan, Rogerson and Wright (1997). The logic is straightforward: by adding a vector of serially and contemporaneously uncorrelated measurement errors, we complete the probability space of the model (the theoretical covariance matrix of $[y_{1t}, y_{2t}]$ is no longer singular). Since actual variables typically fail to match their model counterparts (e.g. actual savings are typically different from model based measures of savings), the addition of measurement errors is easily justifiable. Note that, if this route is taken, a simple diagnostic on the quality of the model can be obtained by comparing the size of the estimated standard deviation of the measurement errors and of the structural shocks. Standard deviations for the former much larger than for the latters suggest that misspecification is likely to be present.

Example 6.18 *In example 6.15, if we wish to complete the probability space of the model, we need to add five measurement errors to the vector of shocks. Alternatively, we could use, e.g., real balances and deposits to estimate the parameters of the model. However, it is unlikely that these two series have information to estimate the share of labor in production function* η. *Hence, identification of the parameters may be a problem when using a subset of the variables of the model*

The introduction of a vector of serially and contemporaneously uncorrelated measurement errors does not alter the dynamics of the model. Therefore, the quality of the model's approximation to the data is left unchanged. Ireland (2004), guessing that both dynamic and contemporaneous misspecifications are likely to be present in simple DSGE models, instead adds a VAR(1) vector of measurement errors. The importance of these dynamics for the resulting hybrid model can be used to gauge how far the model is from the data, much in the spirit of Watson (1993), and an analysis of the properties of the estimated VAR may help in respecifying the model (see chapter 7). However, is important to note that the hybrid model can no longer be considered "structural": the additional dynamics play the same role as distributed lags which were added in the past to specifications derived from static economic theory when confronted with the (dynamics of the) data.

The second issue concerns the quality of the model's approximation to the data. It is clear that to estimate the parameters with ML and to validate model, one must assume that it "correctly" represents the process generating the data up to a set of unknown parameters. Some form of misspecification regarding e.g. the distribution of the errors (see White (1982)) or the parametrization (see Hansen and Sargent (1998)), can be handled using the quasi-maximum likelihood approach discussed in section 6.2. However, as we will argue in chapter 7, the misspecification that a DSGE model typically displays is of different type. Adding contemporaneous uncorrelated measurement errors avoids singularities but it does not necessarily reduce misspecification. Moreover, while with GMM one is free to choose the relationships used to estimate the parameters of interest, this is not the case with ML since joint estimation of all the relationships produced by the model is generally performed. Under these conditions, maximum likelihood estimates of the parameters are unlikely to be consistent and economic exercises conducted conditional on these estimates may be meaningless. In other words, credible maximum likelihood estimation of the parameters of a DSGE model requires strong beliefs about the nature of the model.

Third, for parameters to be estimable they need to be identifiable. For example, if θ_1 and θ_2 are parameters and only $\theta_1 + \theta_2$ or $\theta_1\theta_2$ are identifiable, they can not be estimated separately. Besides this generic problem, thoroughly discussed in Hamilton (1994), DSGE models often face partial identifiability problems in the sense that the series used may have little information about the parameters of interest. This is not surprising: estimating, say, parameters of a monetary policy rule out of export or the trade balance is unlikely to be successful even if these parameters appear in the relevant equations. Furthermore, certain parameters affect only the steady state and therefore cannot be estimated when the model is written in deviations from the steady states or when variables are entered in log differences. In this situation two approaches are possible. The first one, which is more standard, is to calibrate nonestimable parameters (say θ_1) and provide ML estimates for the remaining free parameters (say θ_2) conditional on the chosen θ_1. As argued in chapter 7, such a choice may generate consistency problems and distort the asymptotic distribution of θ_2. The alternative is to use other moment conditions were these parameters appear and jointly estimate θ_1 and θ_2 using the scores of the likelihood and these moment conditions. Since the score of the likelihood has the format of moment conditions, this mixed approach will produce, under regularity conditions, consistent and asymptotically normal estimates. When this last alternative is unfeasible, local sensitivity analysis in a neighborhood of the calibrated parameters is advisable to explore the shape of the likelihood function around the maximum for θ_2 one finds.

Note also the similarities between this and the GMM approach described in chapter 5. Two main differences should be noted. First, the construction of the scores requires the solution of the model (or the rational expectation assumption), which was not necessary to estimate parameters with GMM. Second, if no misspecification is present, ML estimates will, by construction, be more efficient than GMM estimates.

Once parameter estimates are obtained one can proceed to validate the model and/or examine the properties of the implied system. Statistical validation can be conducted in

many ways. For example, if interesting economic hypotheses involve restrictions on a subset of the parameters of the model, standard t-tests or likelihood ratio tests using the restricted and the unrestricted versions of the model can be performed.

Example 6.19 *(Money demand equation) Consider a representative agent maximizing* $E_0 \sum_t \beta^t \left[\frac{1}{1-\varphi_c} c_t^{1-\varphi_c} + \frac{\vartheta_M}{1-\varphi_M}\left(\frac{M_{t+1}}{p_t}\right)^{1-\varphi_M}\right]$ *by choice of* (c_t, B_{t+1}, M_{t+1}) *subject to* $c_t + \frac{B_{t+1}}{p_t} + \frac{M_{t+1}}{p_t} + \frac{b_1}{2}\frac{(M_{t+1}-M_t)^2}{p_t} + \frac{b_2}{2}\frac{(M_t - M_{t-1})^2}{p_t} \leq w_t + \frac{M_t}{p_t} + (1+i_t)\frac{B_t}{p_t}$*, where* b_1, b_2 *are parameters,* w_t *is an exogenous labor income and* B_t *are nominal one-period bonds. The two optimality conditions are* $c_t^{-\varphi} = \beta E_t[c_{t+1}^{-\varphi}\frac{p_t}{p_{t+1}}(1+i_{t+1})]$ *and* $\vartheta_M\left(\frac{M_{t+1}}{p_t}\right)^{-\varphi_M} c_t^{\varphi_c} = E_t[1 - \frac{1}{1+i_{t+1}} + (b_1 + \frac{b_2}{1+i_{t+1}})\Delta M_{t+1} - \frac{1}{1+i_{t+1}}(b_1 + \frac{b_2}{1+i_{t+2}})\Delta M_{t+2}]$ *where* $\Delta M_{t+1} = M_{t+1} - M_t$*. Log linearizing the two conditions, solving out for* $\hat{i}_{t+1}$ *and using the budget constraint we have that* $\phi_c \hat{w}_t - \phi_M(\hat{M}_{t+1} - \hat{p}_t) = \alpha_1 \widehat{\Delta M}_{t+1} + \alpha_2 \widehat{\Delta M}_{t+2} + \alpha_3 \widehat{\Delta w}_{t+1} + \alpha_4 \widehat{\Delta w}_{t+2} + \alpha_5 \widehat{\Delta p}_{t+1} + \alpha_6 \widehat{\Delta p}_{t+2}$ *where* α_j *are functions of the deep parameters of the model* (b_1, b_2) *and of the steady states* $i^{ss}, \Delta M^{ss}$*. If we assume that the Central bank chooses* i_{t+1} *so that* $\Delta \hat{p}_t = 0$*, that bonds are in zero net supply, the above equation can be solved for* ΔM_t *as a function of the current and future exogenous labor income* $\hat{w}_t$ *and the current and future levels of real balances* $\hat{M}_{t+1} - \hat{p}_t$*.*

The parameters of this model can be estimated in a number of ways. One is GMM. For example, using as instruments lagged values of money growth, real balances and labor income, one could estimate $(\varphi_M, \varphi_c, b_1, b_2, i^s s, \Delta M^{ss}, \beta)$ *from the above equation. Alternatively, one could use ML. To do so the above equation needs to solved forward in order to express current growth rate of money as a function of current and future consumption and current money holdings. As we will see in example 6.20, this is easier to do if we represent the available data with a VAR.*

Since there is only one shock (the exogenous labor income) and the system of equations determining the solution is singular. There are three alternatives to deal with this problem. The one we have used expresses the solution of ΔM_t *in terms of current and future labor income and real balances. Then estimates of the parameters can be found maximizing the likelihood of the resulting equation. The second is to attach to the policy equation an error,* $\Delta \hat{p}_t = \epsilon_{3t}$*. This is easily justifiable if inflation targeting is only pursued on average over some period of time. The third is to assume that labor income is measured with error. In the latter two alternatives, the joint likelihood function of the money demand equation and of the consumption Euler equation can be used to find estimates of the parameters. Note also that not all the parameters may be identifiable from the first setup - the forward looking solution requires elimination of the unstable roots which may have important information about, e.g., the adjustment cost parameters.*

Restricted and unrestricted specifications can also be compared in an out-of-sample forecasting race; for example, using the MSE of the forecasts, or the record of turning point predictions.

Exercise 6.21 *Consider two versions of a RBC model, one with capacity utilization and one without. Describe a Monte Carlo procedure to verify which model matches turning*

points of US output growth better. How would you compare models which are not nested (say, one with capacity utilization and one with adjustment costs to capital)?

The stability of the estimates over subsamples can be examined in a standard way. For example, one can split the sample in two and construct a distance test of the form $\mathsf{S} = (\theta^1 - \theta^2)(\Sigma_{\theta^1} + \Sigma_{\theta^2})^{-1}(\theta^1 - \theta^2)$ where θ^1 is the ML estimate obtained in the first sample and Σ_{θ^1} its estimated covariance matrix and θ^2 is the ML estimate obtained in the second sample and Σ_{θ^2} the corresponding estimated covariance matrix. Recursive tests of this type can also be used to determine when a structural break occurs. That is, for each $1 < \tau < T$, we can construct S_τ by estimating the model over two samples $[1, \tau], [\tau + 1, T]$. Then one would compare $\sup_\tau \mathsf{S}_\tau$ to a $\chi^2(\dim(\theta))$, much in the same spirit as structural stability tests described in chapter 4.

We have seen that the solution of DSGE models can be alternatively written in a state space or restricted VAR(1) format. This latter offers an alternative framework to compare the model to the data. The restrictions that DSGE imposes on VARs are of two types. First, log-linearized DSGE models are typically VAR(1) models. Therefore, the methods described in chapter 4 can be used to examine whether the actual data can be modelled as an VAR(1). Second, it is well known, at least since Sargent (1979), that rational expectations models impose an extensive set of cross equations restrictions on the VAR of the data. These restrictions can be used to identify and estimate the free parameters and to test the validity of the model. We discuss how this can be done next.

Example 6.20 *(Kurmann) Consider an hybrid Phillips curve, $\pi_t = \alpha_1 E_t \pi_{t+1} + \alpha_2 \pi_{t-1} + \alpha_3 mc_t + e_t$ which can be obtained from a standard sticky price model once a fraction of the producers fix the price using a rule of thumb and adding some measurement error e_t. The rule necessary to produce such an expression is that the new price is set to an average of last period's price, updated with last period's inflation rate (as in Gali and Gertler (1999)). Assume mc_t is exogenous and let $\mathcal{F}_t$ represents the information set available at each t. For any $z_t \in \mathcal{F}_t$, $E_t(E_t[y_{t+\tau}|\mathcal{F}_t]|z_t) = E_t(y_{t+\tau}|z_t)$, by the law of iterated expectations. Let $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$ be the companion form representation of the model where $\mathbb{Y}_t$ is of dimension $mq \times 1$ (m variables with q lags each). Since $E_t(mc_{t+\tau}|\mathbb{Y}_t) = \mathcal{S}_1 \mathbb{A}^\tau \mathbb{Y}_t$ and $E(\pi_{t+\tau}|\mathbb{Y}_t) = \mathcal{S}_2 \mathbb{A}^\tau \mathbb{Y}_t$ where $\mathcal{S}_1$ and $\mathcal{S}_2$ are selection matrices, a hybrid Phillips curve implies $\mathcal{S}_2[\mathbb{A} - \alpha_1 \mathbb{A}^2 - \alpha_2 I] = \alpha_3 \mathcal{S}_1 \mathbb{A}$ which produce mq restrictions. For example, if $q = 1$, $\mathbb{Y}_t$ includes only the labor share and inflation and $A_{ii'}$ are the parameters of the VAR we have*

$$\begin{aligned} A_{12} - \alpha_1 A_{12} A_{11} - \alpha_1 A_{22} A_{12} - \alpha_2 &= \alpha_3 A_{11} \\ A_{22} - \alpha_1 A_{21} A_{12} - \alpha_1 A_{22}^2 - \alpha_2 &= \alpha_3 A_{21} \end{aligned} \tag{6.27}$$

(6.27) requires that expectations of real marginal costs and inflation produced by a VAR are consistent with the dynamics of the model. One way to impose these restrictions is to express the coefficients of the inflation equation in the VAR as a function of the remaining $(m-1)mq$ VAR coefficients and the parameters of the theory. Here, since there are four

unknowns and two equations, the system can be solved for, e.g., A_{21} and A_{22} as a function of A_{11} and A_{12}. The likelihood function for the restricted VAR system can then be constructed using the prediction error decomposition and tests of the restrictions obtained comparing the likelihood of restricted and unrestricted VARs.

Exercise 6.22 *Consider an endowment economy where agents receive a random income y_t and may either consume or save it. Suppose that stocks S_{t+1} are the only asset, that their price is p_t^s and that the budget constraint is $c_t + p_t^s S_{t+1} = y_t + (p_t^s + sd_t)S_t$ where sd_t are dividends. Assume $u(c) = \frac{c^{1-\varphi}}{1-\varphi}$ and that agents discount the future at the rate β.*
i) Derive a log-linearized expression for the price of stocks as a function of future dividends, future prices and current and future consumption.
ii) Assume that data on stock prices and stock dividends are available and that an econometrician specifies the process for the data as a VAR of order 2. Derive the cross-equation restrictions that the model imposes on the bivariate representation of prices and dividends (Hint: use the equilibrium conditions to express consumption as a function of dividends).
iii) Assume that also data on consumption is available. Does your answer in ii) change?

Exercise 6.23 *Continuing with example 6.19, consider the log-linearized money demand equation alone. Assume that $\Delta X_t = [\Delta p_t, \Delta w_t, \Delta i_t]$ is exogenous and follows a VAR(q) model which we write in the companion form $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t$ and $\Delta X_t = \mathbb{S}\mathbb{Y}_t$, where $\mathbb{S}$ is a selection matrix. Show that the forward solution can be written as $\Delta \ln M_t = \frac{M_0}{1-\lambda_1} - (1 - \frac{1+i^{ss}}{\lambda_1}(\ln M_{t-1} - \tilde{\phi}'\mathbb{S}\mathbb{Y}_{t-1}) + \frac{\lambda_1 - 1 - (1+i^{ss})}{\lambda_1 - 1}\tilde{\phi}\mathbb{S}(1 - \frac{\mathbb{A}}{\lambda_1})^{-1}\mathbb{Y}_t + v_t$ where $\phi = [1, \frac{\phi_c}{\phi_M}, -\frac{1}{i^s s \phi_M}]$, $\tilde{\phi} = \frac{i^{ss}\varphi_M}{M^{ss}(b_1 + b_2/(1+i^{ss}))(\lambda_1 - (1+i^{ss}-1)}\phi$, λ_1 is the stable solution of $\lambda^2 - (1 + (1+i^{ss}) + (\frac{i^{ss}\varphi_M}{M(b_1 + b_2/(1+i^s s)}) \lambda + 1 = 0$, and v_t is a measurement error, which appears because the econometrician information may be different from the one of the agents. Give the structure of v_t. Show the format of the solution when $q = 2$ and there is no constant in the VAR for ΔX_t. What parameters can you estimate? Write down the likelihood function you want to maximize and show the implied cross equations restrictions.*

Statistical validation is usually insufficient for economic purposes, since it offers scarce indications on the reasons why the model fails and provides very little information about the properties of the estimated model. Therefore, as we have done in chapter 5, one would also like to compare the predictions of the model for a set of interesting statistics of the data. Several statistics can be used. For example, given ML estimates, one could compute unconditional moments such as variability, cross correlations, spectra or cross spectra and compared them with those in the data. To learn about the dynamic properties of the estimated model one could compute impulse responses, variance and historical decompositions. Informal comparisons are typically performed but there is no reason to do so, especially in a ML context. In fact, since $\sqrt{T}(\theta_{ML} - \theta_0) \xrightarrow{D} \mathbb{N}(0, \Sigma_\theta)$, we can compute the asymptotic distribution of any continuous function of θ using the δ-method i.e. if $h(\theta)$ is continuously differentiable, $\sqrt{T}(h(\theta_{ML}) - h(\theta_0)) \xrightarrow{D} N(0, \Sigma_h = \frac{\partial h(\theta)}{\partial \theta}\Sigma_\theta \frac{\partial h(\theta)}{\partial \theta}')$. If an estimate h_T is available in the data, a formal measure of distance between the model and the

data is $(h(\theta_{ML}) - h_T)(\Sigma_h + \Sigma_{h_T})^{-1}(h(\theta_{ML}) - h_T)$, which is asymptotically distributed as a $\chi^2(dim(h))$. Small sample versions of such tests are also easily designed.

Exercise 6.24 *Suppose that* $\sqrt{T}(\theta_{ML} - \theta_0) \xrightarrow{D} N(0, \Sigma_\theta)$ *and suppose that, for the statistic* $h(\theta)$ *of interest, both* h_T *and its standard error are available. Describe how to perform a small sample test of the fit of the model.*

Once the model is found to be adequate in capturing the statistical and the economic features of the data, welfare measures can be calculated and policy exercises performed.

Exercise 6.25 *(Blanchard and Quah) The model described in section 3.1 of chapter 3 produces a solution of the form*

$$\begin{aligned} \Delta GDP_t &= \epsilon_{3t} - \epsilon_{3t-1} + (1+a)\epsilon_{1t} - a\epsilon_{1t-1} \\ UN_t &= -\epsilon_{3t} - a\epsilon_{1t} \end{aligned} \tag{6.28}$$

where $\Delta GDP_t = GDP_t - GDP_{t-1}$ *and* $UN_t = N_t - N^{fe}$ *where* N^{fe} *is full employment equilibrium,* ϵ_{1t} *is a technology shock and* ϵ_{3t} *a money shock.*
i) Transform (6.28) into a state space model
ii) Using data for output growth and appropriately detrended unemployment provide a maximum likelihood estimate of α *and test three hypotheses,* $\alpha = 0$ *and* $\alpha \pm 1$.
iii) Provide impulse responses to technology and money shocks using α_{ML}. *Compare them with those obtained with a structural VAR identified using long run restrictions.*

Exercise 6.26 *(Habit persistence) Consider a basic RBC model driven by technology disturbances and three separate specifications for preferences. The first one assumes intertemporal separability of consumption and leisure, that is* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{c_t^{1-\varphi}}{1-\varphi} + \ln(1-N_t)$. *The second that there is habit persistence in consumption, that is* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{(c_t+\gamma c_{t-1})^{1-\varphi}}{1-\varphi} + \ln(1-N_t)$. *The third that there is habit persistence in leisure so that* $u(c_t, c_{t-1}, N_t, N_{t-1}) = \frac{c_t^{1-\varphi}}{1-\varphi} + \ln(1-N_t+\gamma(1-N_{t-1}))$. *The resource constraint is* $c_t + K_{t+1} = \zeta_t K_t^{1-\eta} N_t^{\eta} + (1-\delta)K_t$ *where* $\ln \zeta_t$ *is an AR(1) process. Using US data on consumption, hours, output and investment estimate the free parameters of the three models assuming that consumption, investment and output are measured with error and that each of these errors is a contemporaneously uncorrelated martingale difference process. Test the hypotheses that habit persistence either in consumption or in leisure is unnecessary to match the data. Compare the responses of the three models to technology shocks. What is the role of habit persistence in propagating technology disturbances? (Hint: Nest the three models in one general specification and test the restrictions).*

Exercise 6.27 *(Woodford) Suppose agents maximize* $E_0 \sum_t \beta^t \epsilon_{4t}[u_1(c_t + G_t) + u_2(\frac{M_t}{p_t}) - \epsilon_{2t} u_3(N_t)]$ *where* G_t *is government expenditure,* $\frac{M_t}{p_t}$ *are real balances,* N_t *is hours,* $c_t = (\int c_{it}^{\frac{1}{\varsigma_p+1}} di)^{\varsigma_p+1}$ *and* $p_t = (\int p_{it}^{-\frac{1}{\varsigma_p}} dj)^{-\varsigma_p}$. *Here* ϵ_{4t} *is a aggregate demand shock and* ϵ_{2t}

a labor supply shock and ς_p the elasticity of substitution across consumption goods. Let aggregate demand for good i be $c_{it} = c_t(\frac{p_{it}}{p_t})^{-\frac{1+\varsigma_p}{\varsigma_p}}$. The budget constraint of consumers is $c_t + \frac{M_t}{p_t} + \frac{B_t}{p_t} + T_t = w_t N_t + \frac{M_{t-1}}{p_t} + \frac{(1+i_{t-1})}{\pi_t}\frac{B_{t-1}}{p_{t-1}}$, where $\frac{B_t}{p_t}$ are real bonds and π_t the inflation rate. Suppose $c_{it} = N_{it}$ and that the price index evolves according to $p_t = (\zeta_p p_{t-1}^{-\frac{1}{\varsigma_p}} + (1-\zeta_p)\tilde{p}_t^{-\frac{1}{\varsigma_p}})^{-\varsigma_p}$, where $\tilde{p}_t$ is the optimal price in a Calvo style setting and ζ_p the fraction of firms not changing prices. Finally, assume that the monetary authority sets interest rates according to $1 + i_t = a_1 + a_2\pi_t + (1+i^{ss})M_t^g$, where i^{ss} is the steady state net interest rate and the fiscal authorities sets T_t according to $T_t = a_3 + a_4\frac{B_{t-1}}{p_{t-1}} + T^{ss}T_t^g$ where T^{ss} are steady state lump sum taxes.
i) Derive the log linearized first order conditions (around the steady states) of the model.
ii) Derive a state space representation for the conditions in i) in terms of $\hat{\epsilon}_t = [\hat{\epsilon}_{4t}, \hat{\epsilon}_{2t}, \hat{M}_t^g, \hat{T}_t^g]$.
iii) Assuming that $\hat{\epsilon}_t$ is an AR(1) with diagonal persistence matrix and that output and inflation are measured with error, provide ML estimates of the parameters of the model using US data for debt, real balances, inflation, output, nominal interest rate and real deficit. Test the hypothesis $a_4 < \frac{1-\beta}{\beta}$ and $a_2 < \frac{1}{\beta}$, which corresponds to passive fiscal policy and active monetary policy in the terminology of Leeper (1991). What is the effect of shocks to T_t^g in the economy?

6.5 Estimating a sticky price model: an example

The model we consider is the same as in exercise 3.2 of chapter 3. Our task is to estimate its structural parameters, test interesting economic hypotheses concerning the magnitude of the coefficients, compare the forecasting performance relative to an unrestricted VAR and, finally, compare some conditional moment implications of the model and of the data.

For convenience we repeat the basic setup: the representative household maximizes $E_0\sum_t \beta^t[\ln c_t + \vartheta_M \ln(\frac{M_t}{p_t}) - \frac{\vartheta_N}{1-\varphi_n}N_t^{1-\varphi_n} - \frac{\vartheta_{ef}}{1-\varphi_{ef}}Ef_t^{1-\varphi_{ef}}]$ where $c_t = (\int c_{it}^{\frac{1}{\varsigma_p+1}}di)^{\varsigma_p+1}$ is aggregate consumption, ς_p is the elasticity of substitution among consumption goods, $p_t = (\int p_{it}^{-\frac{1}{\varsigma_p}}dj)^{-\varsigma_p}$ is the aggregate price index, $\frac{M_t}{p_t}$ are real balances, N_t is hours worked and Ef_t is effort. The budget constraint is $\int_0^1 p_{it}c_{it}di + M_t = W_{Nt}N_t + W_{et}Ef_t + M_{t-1} + T_t + Prf_t$ where T_t are monetary transfers, Prf_t profits distributed by the firms and W_{Nt}, W_{et} are the reward to working and to effort. A continuum of firms produce differentiated good using $c_{it} = \zeta_t(N_{it}^{\eta_2}Ef_{it}^{1-\eta_2})^{\eta_1}$ where $N_{it}^{\eta_2}Ef_{it}^{1-\eta_2}$ is the quantity of effective input and ζ_t an aggregate non-stationary technology shock, $\Delta\zeta_t = \epsilon_{1t}$ where $\ln\epsilon_{1t} \sim iid\ \mathbb{N}(0, \sigma_\zeta^2)$. Firms set prices one period in advance, taking as given the aggregate price level and not knowing the current realization of the shocks. Once shocks are realized, firms optimally choose employment and effort. So long as marginal costs are below the predetermined price, firms will meet the demand for their product and choose an output level equal to $c_{it} = (\frac{p_{it}}{p_t})^{-1-\varsigma_p^{-1}}c_t$. Optimal price setting implies $E_{t-1}[\frac{1}{c_t}((\eta_1\eta_2)p_{it}c_{it} - (\varsigma_p+1)W_{Nt}N_{it})] = 0$ which, in the absence of uncertainty, reduces to the standard that condition that the price

is a markup over marginal costs. We assume that the monetary authority controls the quantity of money and sets $\Delta M_t = \epsilon_{3t} + a_M \epsilon_{1t}$ where $\ln \epsilon_{3t} \sim iid\ \mathbb{N}\ (0, \sigma_M^2)$ and a_M is a parameter. Letting lower case letters denote natural logs, the model implies the following equilibrium conditions for inflation (Δp_t), output growth (Δgdp), employment (n_t) and labor productivity growth (Δnp_t)

$$\begin{aligned}
\Delta p_t &= \epsilon_{3t-1} - (1 - a_M)\epsilon_{1t-1} && (6.29)\\
\Delta gdp_t &= \Delta\epsilon_{3t} + a_M\epsilon_{1t} + (1 - a_M)\epsilon_{1t-1} && (6.30)\\
n_t &= \frac{1}{\eta}\epsilon_{3t} - \frac{1 - a_M}{\eta}\epsilon_{1t} && (6.31)\\
\Delta np_t &= (1 - \frac{1}{\eta})\Delta\epsilon_{3t} + (\frac{1 - a_M}{\eta} + a_M)\epsilon_{1t} + (1 - a_M)(1 - \frac{1}{\eta})\epsilon_{1t-1} && (6.32)
\end{aligned}$$

where $np_t = gdp_t - n_t$ and $\eta = \eta_1(\eta_2 + (1 - \eta_2)\frac{1+\varphi_n}{1+\varphi_{ef}})$.

The model therefore has two shocks (a technology and a monetary one) and implications for at least four variables ($\Delta p_t, \Delta gdp_t, \Delta np_t, n_t$). There are 11 free parameters ($\eta_1, \eta_2, \varphi_n, \varphi_{ef}, \beta, \sigma_\zeta^2, \sigma_M^2, a_M\ \vartheta_M, \vartheta_n, \vartheta_{ef}$), but many of them are do not appear in or are not identifiable from (6.29)-(6.32). In fact it is easy to verify that only a_M and η independently enter the four conditions and therefore, together with σ_ζ^2 and σ_M^2, are the only ones estimable with likelihood methods.

Since there are only two shocks the covariance matrix produced by the model is singular and we are free to choose which two variables to use to estimate the parameters. In the baseline case we select productivity and hours. As a robustness check, we repeat estimation using both output and hours, and prices and output. Note that, in this latter case, also η is non-identifiable. As an alternative, we estimate the model adding serially uncorrelated measurement errors to output and productivity. In this case we estimate six parameters: the four structural ones and the variances of the two measurement errors.

We examine both the statistical and economic fit of the model. First, we study several specifications which restrict a_M and/or η to some prespecified value. A Likelihood ratio test is performed in each case and the statistics compared to a χ^2 distribution. For the specification with measurement errors, we also perform a forecasting exercise comparing the one step ahead MSE of the model to the MSE produced by a four variable VAR(1) model, which has 20 parameters (four constants and 16 autoregressive coefficients). Since the number of coefficients in the two specifications differs, we also compare the two specifications with a Schwarz criterion (see chapter 4). In this latter case, the VAR model is penalized since it has a larger number of parameters. We also compute tests of forecasting accuracy, as detailed in section 6.2. Conditional on the estimated parameters, we compute impulse responses, to examine the sign of the dynamics of the variables to technology and monetary shocks, and compare few elements of the unconditional autocovariance function for the four variables in the model and in the data.

We use CPI, GDP (constant in 1992 prices) and total hours (equal to average weekly hours multiplied by civilian employment) for Canada for the period 1981:2-2002:3. All

variables are logged and first differences of the log are used to compute growth rates. Total hours are detrended using a linear trend.

(6.29)-(6.32) has a state space representation for $\alpha = [\epsilon_{1t}, \epsilon_{1t-1}, \epsilon_{3t}, \epsilon_{3t-1}, v_{1t}, v_{2t}]$, where $v_{it}, i = 1, 2$ are measurement errors, $x_{1t} = \begin{bmatrix} 0 & a_M - 1 & 0 & 1 & 0 & 0 \\ a_M & 1 - a_M & 1 & -1 & 1 & 0 \\ \frac{a_M - 1}{\eta} & 0 & \frac{1}{\eta} & 0 & 0 & 0 \\ \frac{1-a_M}{\eta} + a_M & \frac{(1-a_M)(\eta-1)}{\eta} & \frac{\eta-1}{\eta} & -\frac{\eta-1}{\eta} & 0 & 1 \end{bmatrix}$,

$\mathbb{D}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$, $\mathbb{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\Sigma_{v_1} = 0$ with the appropriate adjustments if no measurement error is included. The Kalman filter is initialized using $\alpha_{1|0} = 0$ and $\Omega_{1|0} = I$. The likelihood function is computed recursively and a simplex method is used to locate the maximum. We use this approach instead of a one based on the gradient because the likelihood is flat, the maximum is around the boundary of the parameter space and convergence is hard to achieve. The cost is that no standard errors for the estimates are available. Table 6.1 reports parameter estimates, together with the p-values of various likelihood ratio tests.

Table 6.1: ML estimates

Data set	a_M	η	σ^2_ζ	σ^2_M			Log Likelihood
$(\Delta np_t, n_t)$	0.5533	0.9998	1.06e-4	6.69e-4			704.00
$(\Delta gdp_t, n_t)$	-7.7336	0.7440	6.22e-6	1.05e-4			752.16
$(\Delta gdp_t, \Delta p_t)$	3.2007		1.26e-5	1.57e-4			847.12
	a_M	η	σ^2_ζ	σ^2_M	σ^2_{v1}	σ^2_{v2}	Log Likelihood
$(n_t, \Delta np_t, \Delta gdp_t, \Delta p_t)$	-0.9041	1.2423	5.82e-6	4.82e-6	0.0236	0.0072	1336
Restrictions	$a_M = 0$	$\eta = 1$	$\eta = 1$	$\eta = 1.2$			
	$a_M = -1.0$						
$(\Delta np_t, n_t)$, p-value	0.03	0.97	0.01	0.00			
$(\Delta gdp_t, n_t)$, p-value	0.00	0.00	0.00	0.00			
$(n_t, \Delta np_t, \Delta gdp_t, \Delta p_t)$ p-value	0.00	0.001	0.00	0.87			
Restrictions	$a_M = 0$	$a_M = 1$	$a_M = -1.0$				
$(\Delta y_t, \Delta p_t)$, p-value	0.00	0.00	0.00				

Several features of the table deserve comments. First, using bivariate specifications the estimated value of η is less then one. Since for $\varphi_{ef} = \varphi_N$, $\eta = \eta_1$, this implies that there is no evidence of short run increasing returns to scale. The lack of increasing returns is formally confirmed by likelihood ratio tests: conditioning on values of $\eta \geq 1$ reduces the likelihood. However, when measurement errors are included, mild short run increasing returns to scale obtain. Second, the estimated value of a_M depends on the data set: it is positive and moderate when productivity and hours are used; positive and large when output and prices

are used, strongly negative when output and hours are used and moderately negative when the four series are used. The reason for this large variety of estimates is that the likelihood function is very flat in the a_M dimension. Figure 6.1 illustrates this fact using the first data set. It is easy to see that $a_M = 0$, or $a_M = -0.5$ are not extremely unlikely.

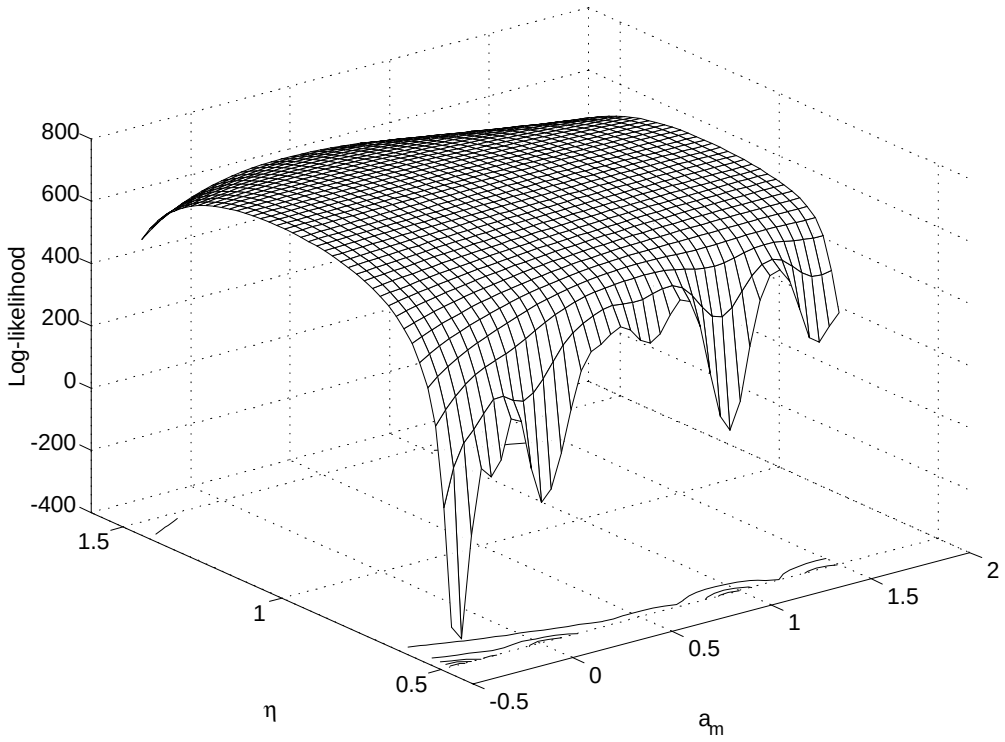

Figure 6.1: Likelihood surface

Note that, at face value, these estimates imply that monetary policy is countercyclical in two specifications and mildly accommodative in the two others. Third, the likelihood function is also relatively flat in the $\sigma_\zeta^2, \sigma_M^2$ space and achieves the maximum around the boundary of the parameter space. Note that with all bivariate data sets, and somewhat counterintuitively, the variance of monetary shocks is estimated to be larger than the variance of the technology shocks. Fourth, the size of the estimated variance of measurement errors is several orders of magnitude larger than the estimated variance of structural shocks, suggesting that misspecification is likely to be present.

Forecasts produced by the model are poor. In fact, the one-step ahead MSEs for hours, productivity growth, output growth and inflation are 30, 12, 7, 15 times larger than the

ones produced by a VAR(1). A test for forecasting accuracy confirms that the forecasts of the model are different from those produced by a VAR(1). The picture improves when a penalty for the larger number of parameters is used. In this case, the value of the Schwartz criterion for the model is "only" twice as large as the one of the VAR(1).

Impulse responses to unitary positive technology and money shocks are in figure 6.2. We report responses obtained with the parameters estimated using productivity and hours data (DATA 1) and output and hours data (DATA 2). Several features are worth discussing. First, estimates of a_M do not affect the responses to monetary shocks. Second, qualitatively speaking, and excluding the responses of output to technology disturbances, the dynamics induced by the shocks are similar across parametrizations. Third, the shape of the responses to technology and monetary shocks looks very similar (up to a sign change) when productivity and hours data are used. Hence, it would be hard to distinguish the two type of shocks by looking at the comovements of these two variables only. Fourth, as expected, the response of productivity to technology shocks is permanent (there is an initial overshooting) and the response of hours is temporarily negative.

Table 6.2 reports cross covariances in the model and in the data. A few features of the table stand out. First, the model estimated with measurement errors fails to capture, both quantitatively and qualitatively, the cross covariance of the data: the magnitude of the estimated covariances is 10 times smaller than the one in the data and the signs of the contemporaneous covariance of $(n_t, \Delta np_t)$, $(\Delta gdp_t, \Delta p_t)$ and $(\Delta np_t, \Delta np_{t-1})$ are wrong.

Second, cross covariances obtained when the model is estimated using productivity and hours data are still somewhat poor. For example, the estimated covariances of $(\Delta gdp_t, \Delta gdp_{t-1})$ and $(\Delta np_t, \Delta np_{t-1})$ are ten times larger than in the data and a distance test rejects the hypothesis that the two set of cross covariances are indistinguishable. Despite these failures, the model estimated using hours and productivity data, captures two important qualitative features of the data: the negative contemporaneous covariance between hours and productivity and the negative lagged covariance of productivity.

Finally, note that neither of the two specifications can reproduce the negative covariance between output growth and inflation found in the data.

Moments/Data	$(\Delta np_t, n_t)$	$(\Delta np_t, n_t, \Delta p_t, \Delta gdp_t)$	Actual data
$\text{cov}(\Delta gdp_t, n_t)$	6.96e-04	4.00e-06	1.07e-05
$\text{cov}(\Delta gdp_t, \Delta np_t)$	5.86e-05	1.56e-06	1.36e-05
$\text{cov}(\Delta np_t, n_t)$	-4.77e-05	1.80e-06	-4.95e-05
$\text{cov}(\Delta gdp_t, \Delta p_t)$	6.48e-04	2.67e-06	-2.48e-05
$\text{cov}(\Delta gdp_t, \Delta gdp_{t-1})$	6.91e-04	3.80e-06	3.443-05
$\text{cov}(\Delta np_t, \Delta np_{t-1})$	-1.51e-04	1.07e-06	-2.41e-05

Table 6.1: Cross covariances

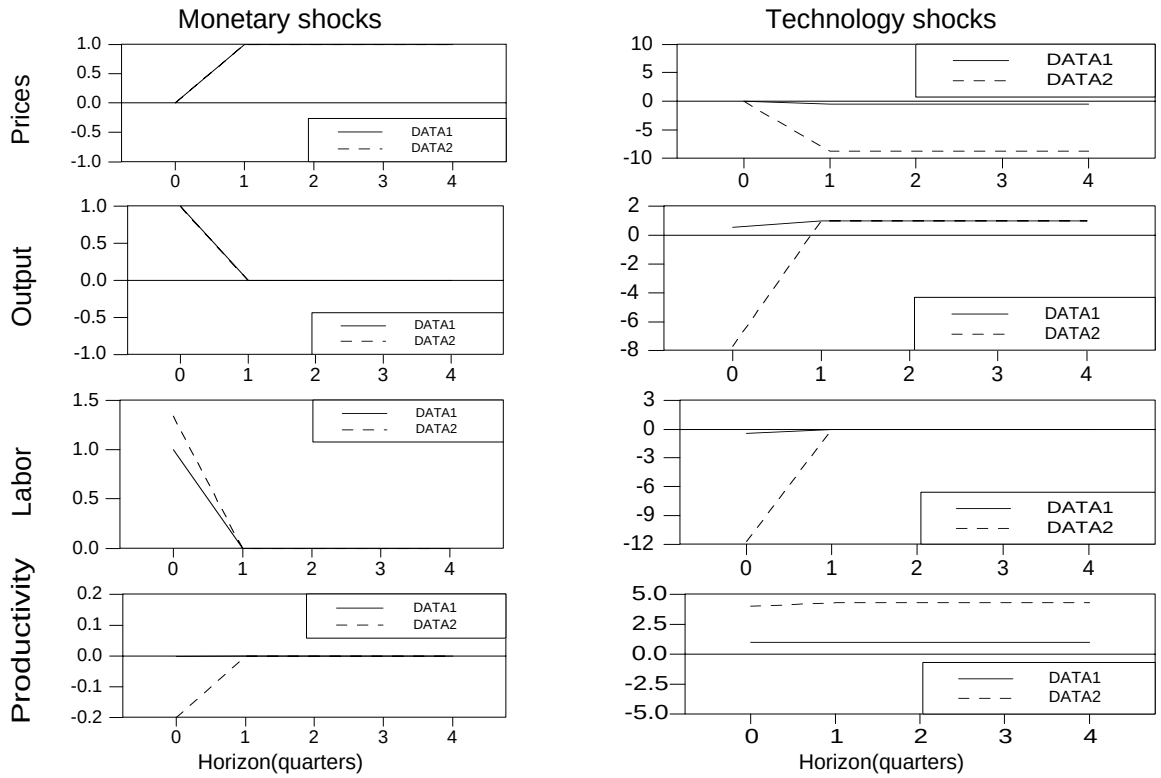

Figure 6.2: Impulse responses