Likelihood methods for DSGE models

Fabio Canova
EUI and CEPR
January 2014

Outline

- State space models and the Kalman Filter.
- Prediction error decomposition of the likelihood.
- Numerical routines.
- ML estimation of DSGE models.
- Three examples.

References

Altug, S. (1989). "Time to build and Aggregate Fluctuations: Some New Evidence", *International Economic Review*, 30, 883-920.

Andreasan, M. (2010). How to maximize the likelihood function for a DSGE model, *Computational Economics*, 35, 127-154.

Canova, F. and Menz, T., 2011, Does Money have a Role in Shaping Domestic Business Cycles. An International Investigation (with T. Menz), *Journal of Money Credit and Banking*, 43(4), 577-609, 2011.

Christiano, L. and Vigfusson, R. 2003. Maximum likelihood in freuqncy domain: the importance of time to plan. *Journal of Monetary Economics* 50, 789-815.

Gali, J. (1999) Technology, Employment and Business Cycle: Do Technology Shocks Explain Aggregate Fluctuations? *American Economic Review*, 89, 249-271.

Harvey, C. (1991), *Time Series Models*, Halstead Press.

Hansen, L and Sargent, T. (1993) Seasonality and approximation errors in rational expectations models. *Journal of Econometrics*, 55,21-55.

Hamilton, J. (1994), *Time Series Models*, Princeton University Press.

Hansen, L. and Sargent, T.(1998), *Recursive linear Models of Dynamic Economies*, University of Chicago, manuscript.

Kim, J. (2000) Constructing and Estimating a realistic Optimizing Model of Monetary Policy, *Journal of Monetary Economics*, 45, 329-359

Ireland, P. (2004) A method for taking Models to the data, *Journal of Economic Dynamics and Control*, 28, 1205-1226.

Leeper, E. and Sims, C. (1994), " Toward a Modern Macroeconomic Model Usable for Policy Analyses", In Rotemberg, J. and Fisher, S. (eds.) *NBER Macroeconomic Annual*, 9, MIT Press.

Ljung, L. and Soderstroem, T. (1983) *Theory and Practice of Recursive Identification*, Cambridge, MIT Press.

Kurmann, A. (2003) ML estimation of Dynamic Stochastic Theories with an Application to New Keynesian Pricing, University of Quebec at Montreal.

Watson, M. (1989) Recursive Solution methods for Dynamic Linear Rational Expectations Models, *Journal of Econometrics*, 41, 65-89.

White, H. (1982) Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50, 1-25.

1 Why full information ML and not GMM (SMM)?

• GMM typically used in limited information settings (only a few equations of the model considered).

• Important small sample distortions (especially in estimate of weighting matrix, J-tests).

• GMM and SMM not designed for model comparison.

• Fail to satisfy likelihood principle (all information in an experiment is contained in the likelihood of parameters).

2 State Space Models

$$
\begin{aligned}
y_t &= x'_{1t}\alpha_t + x'_{2t}v_{1t} && v_{1t} \sim iid\ \mathbb{N}(0, \Sigma_{v_1}) && (1)\\
\alpha_t &= \mathbb{D}_{0t} + \mathbb{D}_{1t}\alpha_{t-1} + \mathbb{D}_{2t}v_{2t} && v_{2t} \sim iid\ \mathbb{N}(0, \Sigma_{v_2}) && (2)
\end{aligned}
$$

x'_{1t} is $m \times k_1$, x'_{2t} is $m \times k_2$; $\mathbb{D}_{0t}$ is $k_1 \times 1$, $\mathbb{D}_{1t}$ is $k_1 \times k_1$, $\mathbb{D}_{2t}$ is $k_1 \times k_3$,.
Assume: $E(v_{1t}v'_{2\tau}) = 0$ and $E(v_{1t}\alpha'_0) = 0\ \forall t, \tau$.

(1) is a measurement equation and y_t are the observables. (2) is a transition (state) equation and α_t is an unobserved state.

Why normality of the errors? If $\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N(\begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix})$, then $z_1|z_2 \sim N((\bar{z}_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \bar{z}_2)); (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})]$.

State space framework is very general. Many popular time series models and many interesting economic models with latent variables fit (1)-(2).

i) Any ARMA(q_1, q_2) fits (1)-(2).

Example 1 $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t + D_1 e_{t-1}$ *can be written as:*

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

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

which fits (1)-(2) for $\alpha_t = \begin{bmatrix} y_t \\ A_2 y_{t-1} + D_1 e_t \end{bmatrix}$, $\mathbb{D}_{1t} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix}$,

$\mathbb{D}_{2t} = \begin{bmatrix} 1 \\ D_1 \end{bmatrix}$, $\mathbb{D}_{0t} = 0$, $x'_{1t} = [1,\ 0]$, $\Sigma_{v_1} = 0$, $\Sigma_{v_2} = \sigma_e^2$.

ii) Any VAR(p) fits (1)-(2) (with or without TVC)

Example 2 $y_t = A(\ell)y_{t-1} + e_t$. *Use companion form representation* $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + \mathbb{E}_t$ *where* $\mathbb{E}_t = [e_t, 0, \ldots 0]'$ *and*

$$\mathbb{A} = \begin{bmatrix} A_1 & A_2 & \ldots & \ldots & A_q \\ I & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix}$$

Trivially fits (1)-(2) for $x'_{1t} = [I, 0, \ldots 0], \alpha_t = [y'_t, y'_{t-1}, \ldots, y'_{t-q}],\ \mathbb{D}_{1t} = \mathbb{A},\ \Sigma_{v_1} = 0,\ v_{2t} = \mathbb{E}_t,\ \mathbb{D}_{2t} = I, \mathbb{D}_{0t} = 0$.

Example 3

$$\begin{aligned} y_t &= A_t y_{t-1} + v_{1t} && (3) \\ A_t &= A_{t-1} + v_{2t} && (4) \end{aligned}$$

iii) Any latent variable specifications fits (1) and (2).

Example 4 *Ex-ante real rate of interest:* $\alpha_t \equiv i_t - \pi_t^e = \phi\alpha_{t-1} + v_{2t}$. *Observed real rate:* $y_t \equiv i_t - \pi_t = \alpha_t + v_{1t}$, *where* v_{1t} *is an expectation error.*

Example 5 *Potential output::* $\alpha_t = \rho\alpha_{t-1} + v_{2t}$. *Measured output:* $y_t = \alpha_t + v_{1t}$, v_{1t} *is the output gap.*

Example 6 *Trend/cycle decomposition. Trend:* $\alpha_t = \alpha_{t-1} + v_{2t}$. *Observable data:* $y_t = x_{1t}\alpha_t + x_{2t}v_{1t}$. $x'_{1t} = x'_1$ *is the loadings on the trend and* $x'_{2t} = x'_2$ *the loading on the cycle* v_{1t}

3 Kalman Filter

Can be used to compute optimal forecasts of y_t and recursive estimates of the latent state α_t with time t information for models like (1)-(2).

Let $\alpha_{t|t}$ be the optimal (MSE) estimator of α_t with information up to t; and $\Omega_{t|t}$ the MSE of α_t. Assume $x'_{1t} = x'_1,\ x'_{2t} = x'_2, \mathbb{D}_{1t} = \mathbb{D}_1, \mathbb{D}_{2t} = \mathbb{D}_2, \mathbb{D}_{0t} = \mathbb{D}_0$ are known; $t = 1, \ldots T$ observations.

Five Steps:

• Choose initial conditions. If all eigenvalues of $\mathbb{D}_1$ are less then one in absolute value, $x'_2\Sigma_{v_1}x_2\ D'_2\Sigma_{v_2}D_2$ positive semidefinite and symmetric, 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

$vec(\Omega_{1|0}) = (I - (\mathbb{D}_1 \otimes \mathbb{D}_1')^{-1})vec(\mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2')$, i.e. set initial conditions equal to the unconditional mean and variance of the process. Otherwise, $\alpha_{1|0} = 0$, $\Omega_{1|0} = \kappa * I$, κ large. OK because $\Omega_{1|0}$ symmetric positive semidefinite.

- Forecast y_t and mean square of the forecast error (with $t-1$ info):

$$\begin{aligned} y_{t|t-1} &= x_1'\alpha_{t|t-1} && (5) \\ E(y_t - y_{t|t-1})(y_t - y_{t|t-1})' &= E(x_1'(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'x_1) + x_2'\Sigma_{v_1}x_2 \\ &= x_1'\Omega_{t|t-1}x_1 + x_2'\Sigma_{v_1}x_2 \equiv \Sigma_{t|t-1} && (6) \end{aligned}$$

- Update state estimate: (with t information)

$$\begin{aligned} \alpha_{t|t} &= \alpha_{t|t-1} + \Omega_{t|t-1}x_1'\Sigma_{t|t-1}^{-1}(y_t - x_1\alpha_{t|t-1}) && (7) \\ \Omega_{t|t} &= \Omega_{t|t-1} - \Omega_{t|t-1}x_1\Sigma_{t|t-1}^{-1}x_1'\Omega_{t|t-1} && (8) \end{aligned}$$

Note: $\Omega_{t|t-1}x_1' = E(\alpha_t - \alpha_{t|t-1})(y_t - y_{t|t-1})$.α_t is updated using linear OLS projection of $\alpha_t - \alpha_{t|t-1}$ on $y_t - y_{t|t-1}$ multiplied by the prediction

error. Similarly, $\Omega_{t|t} = E(\alpha_t - \alpha_{t|t-1})(\alpha_t - \alpha_{t|t-1})'$ updated using covariance between forecast errors in the two equations and the MSE of the forecasts.

- Forecast the state next period:

$$\alpha_{t+1|t} = \mathbb{D}_1\alpha_{t|t} + \mathbb{D}_0 = \mathbb{D}_1\alpha_{t|t-1} + \mathbb{D}_0 + \mathfrak{K}_t\epsilon_t \quad (9)$$

$$\Omega_{t+1|t} = \mathbb{D}_1\Omega_{t|t}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' \quad (10)$$

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

- Repeat previous steps until $t = T$.

Note: we use properties of bivariate normal to construct updated mean and variance of α in (7) and (8).

• Smoothing equations: (working backward from y_T), $t = T-1, \ldots, 1$.

$$\alpha_{t|T} = \alpha_{t|t} + (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\alpha_{t+1|T} - \mathbb{D}_1\alpha_{t|t}) \quad (11)$$

$$\Omega_{t|T} = \Omega_{t|t} - (\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})(\Omega_{t+1|T} - \Omega_{t+1|t})(\Omega_{t|t}\mathbb{D}_1'\Omega_{t+1|t}^{-1})' \quad (12)$$

Equations (11)-(12) define the Kalman smoother.

• IMPORTANT: The Kalman smoother is used for extraction of time series components (cycles, trends) not to estimate parameters of a structural model.

Example 7 $y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t$. *Then* $\alpha = [y_t, y_{t-1}]'$, $v_{2t} = [e_t, 0]$, $\mathbb{D}_1 = \begin{bmatrix} A_1 & A_2 \\ 1 & 0 \end{bmatrix}$, $\Sigma_{v_2} = \begin{bmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbb{D}_0 = v_{1t} = 0$, $x_1' = [1, 0]$.

KF Forecasts: $E_{t-1}y_t = A_1 y_{t-1} + A_2 y_{t-2}$; $E(y_t - E_{t-1}y_t)^2 = \sigma_e^2$.

KF Updates: $\alpha_{t|t} = \alpha_{t|t-1} + \Omega_{t|t-1}\sigma_e^{-2}v_{2t}$; $\Omega_{t|t} = \Omega_{t|t-1} + \Omega_{t|t-1}\sigma_e^{-2}\Omega_{t|t-1}$.

• Innovation representation of state space models:

$$\alpha_{t|t-1} = \mathbb{D}_1 \alpha_{t-1|t-1} + \mathbb{D}_0 + \mathfrak{K}_{t-1}\epsilon_t \quad (13)$$

$$y_t = x_{1t}'\alpha_{t|t-1} + \epsilon_t \quad (14)$$

$\epsilon_t =$ one-step ahead forecast error and $E_t(\epsilon_t\epsilon_t') \equiv \Sigma_{t|t-1}$. (13)-(14) is a reduced rank system of equations!

Hansen and Sargent (1998, p. 126-128): $\Omega_{t|t} = \mathbb{D}_1\Omega_{t-1|t-1}\mathbb{D}_1' + \mathbb{D}_2\Sigma_{v_2}\mathbb{D}_2' - \mathbb{D}_1\Omega_{t-1|t-1}x_1(x_1'\Omega_{t-1|t-1}x_1 + x_2'\Sigma_{v_1}x_2)^{-1}x_1'\Omega_{t-1|t-1}\mathbb{D}_1)$ (matrix Riccati equation).

If coefficients are constant, under regularity conditions:

$\lim_{t\to\infty}\Omega_{t|t} = \Omega$; $\lim_{t\to\infty}\mathfrak{K}_t = \mathfrak{K}$ $\lim_{t\to\infty}\Sigma_{t|t} = x_1'\Omega x_1 + x_2'\Sigma_{v_1}x_2 = \Sigma$.

$\Omega, \mathfrak{K}, \Sigma$ are asymptotically equivalent to those obtained with a recursive least square estimator.

Example 8 *GDP potential is* $\alpha_t = \alpha_{t-1}$*; observable GDP is* $y_t = \alpha_t + v_{1t}$*,* v_{1t} iid $\mathbb{N}(0, \sigma^2_{v_1})$*. Then*

$$\Omega_{t|t} = \Omega_{t|t-1} - \Omega_{t|t-1}(\Omega_{t|t-1} + \sigma^2_{v_1})^{-1}\Omega_{t|t-1} = \frac{\Omega_{t|t-1}}{1+\frac{\Omega_{t|t-1}}{\sigma^2_{v_1}}} = \frac{\Omega_{t-1|t-1}}{1+\frac{\Omega_{t-1|t-1}}{\sigma^2_{v_1}}};$$

$$\alpha_{t+1|t+1} = \alpha_{t|t} + \frac{\frac{\bar{\Omega}_0}{\sigma^2_{v_1}}}{1+t\frac{\bar{\Omega}_0}{\sigma^2_{v_1}}}(y_t - \alpha_{t|t}) \text{ and } \lim_{t\to\infty} \alpha_{t+1|t+1} = \alpha_{t|t} = \bar{\alpha}.$$

Hence, when state is a constant, the KF asymptotically produces a constant.

• KF can be applied to models with time varying coefficients, so long as they are linear in parameters e.g.

$$\begin{aligned} y_t &= A_t y_{t-1} + v_{1t} \\ A_t &= A_{t-1} + v_{2t} \end{aligned}$$

• KF can be used in special non-linear state space models e.g. $y_t = \alpha_t + v_{1t}$, $\alpha_{t+1} = \alpha_t\theta + v_{2t}$ and interest is in θ (both θ and α_t are unobservable).

$$\begin{aligned} \alpha_{t|t} &= \theta_{t|t-1}\alpha_{t|t-1} + \mathfrak{K}_{1t}(y_t - \alpha_{t|t-1}) \\ \theta_{t|t} &= \theta_{t|t-1} + \mathfrak{K}_{2t}(y_t - \alpha_{t|t-1}) \end{aligned} \tag{15}$$

$\mathfrak{K}_{1t} = \frac{\theta_{t|t-1}\kappa_{1t} + \alpha_{t|t-1}\kappa_{2t}}{\kappa_{1t} + \sigma^2_{v_1}}, \mathfrak{K}_{2t} = \frac{\kappa_{2t}}{\kappa_{1t} + \sigma^2_{v_1}}$ and κ_{1t} and κ_{2t} involve linear and quadratic terms in $\theta_{t|t-1}$ and $\alpha_{t|t-1}$ and in past Kalman gains (see Liung and Soderstroem (1983), p. 39-40).

• If initial conditions and innovations are normal, Kalman filter **best predictor** (linear or nonlinear) of y_t. Else, it gives **best linear** predictor.

Example 9 *Suppose y_t is driven by a two state Markov process (which switches, e.g. in expansions/recessions), i.e $y_t = a_0 + a_1 s_t + y_{t-1}$. A two state Markov process can be written as*

$$s_t = (1 - p_2) + (p_1 + p_2 - 1)s_{t-1} + v_{1t} \tag{16}$$

where v_{1t} can take values $[1 - p_1, -p_1, -(1 - p_2), p_2]$ with probabilities $[p_1, 1 - p_1, p_2, 1 - p_2]$.

- v_{1t} is non-normal. (it is binomial)

- Corr($v_{1t}, s_{t-\tau}$ $\tau > 0$) = 0, but $v_{1t}, s_{t-\tau}$ are not independent

KF applied to this model is suboptimal: there are other approaches which give forecasts of y_t with smaller MSE.

4 Prediction error decomposition

Basic idea: If $f(y_1, \ldots y_T)$ is the joint density of $y_t,\ t = 1, \ldots T$. Then:

$$\begin{aligned} f(y_1, \ldots y_T) &= f(y_T, |y_{T-1} \ldots y_1) f(y_{T-1}, \ldots y_1) \\ &= f(y_T, |y_{T-1} \ldots y_t) f(y_{T-1}|y_{T-2}, \ldots y_1) f(y_{T-2}, \ldots y_1) \\ &\ldots \\ &= \prod_{j=0}^{J} f(y_{T-j}, |y_{T-j-1} \ldots y_1) f(y_1) \end{aligned} \tag{17}$$

and $\ln f(y_1, \ldots, y_T) \propto \sum_j \ln f(y_{T-j}, |y_{T-j-1} \ldots y_1) + \ln f(y_1)$.

Suppose $y = (y_1, \ldots y_T) \sim \mathbb{N}(\bar{y}, \Sigma_y)$. Let $\phi = (\bar{y}, \Sigma_y)$.

$$\ln \mathcal{L}(y_1, \ldots, y_T|\phi) = -\frac{T}{2} \ln(2\pi) - \frac{\ln |\Sigma_y|)}{2}) - \frac{1}{2}(y - \bar{y})' \Sigma_y^{-1} (y_t - \bar{y}) \tag{18}$$

Brute force approach. Problem: Σ_y is $T \times T$ matrix. Alternative:

Let ln $\mathcal{L}(y_1, \ldots, y_t|\phi) =$ ln $\mathcal{L}(y_1, \ldots, y_{t-1}|\phi)$ ln $\mathcal{L}(y_t|y_{t-1}, \ldots, y_0|\phi)$. Since y_t is normal, both components are normal.

Define:

- $y_{t|t-1}$: best predictor of y_t, given information up to $t-1$.

- $\epsilon_t = y_t - y_{t|t-1} = y_t - E(y_t|y_{t-1}, \ldots y_1) + E(y_t|y_{t-1}, \ldots y_1) - y_{t|t-1}$.

- MSE$=E(\epsilon_t - E(\epsilon_t))^2 = E(y_t - E(y_t|y_{t-1}, \ldots y_1))^2 + E(E(y_t|y_{t-1}, \ldots y_1) - y_{t|t-1})^2$. MSE minimized if $E(y_t|y_{t-1}, \ldots y_1) = y_{t|t-1}$, since first term does not include $y_{t|t-1}$. Then $MSE \equiv \sigma^2_{\epsilon_t} = var(y_t|y_{t-1}, \ldots, y_1)$.

Density of $(y_t|y_{t-1},\ldots)$ for any $t>1$ is:

$$\ln \mathcal{L}(y_t|y_{t-1},\ldots,y_0,\phi) = -\frac{1}{2}\ln(2\pi) - \ln(\sigma_{\epsilon_t}) - \frac{1}{2}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \quad (19)$$

$$\begin{aligned}
\ln \mathcal{L}(y_1,\ldots y_T|\phi) &= \sum_{t=2}^{T} \ln \mathcal{L}(y_t|y_{t-1},\ldots,y_1,\phi) + \ln \mathcal{L}(y_1,\phi) \\
&= -\frac{T-1}{2}\ln(2\pi) - \sum_{t=2}^{t} \ln \sigma_{\epsilon_t} - \frac{1}{2}\sum_{t=2}^{T}\frac{(y_t - y_{t|t-1})^2}{\sigma^2_{\epsilon_t}} \\
&- \frac{1}{2}\ln(2\pi) - \ln \sigma_{\epsilon_1} - \frac{1}{2}\frac{(y_1 - \bar{y}_1)^2}{\sigma^2_{\epsilon_1}} \quad (20)
\end{aligned}$$

• (20) can be computed recursively: it only involves one step ahead prediction errors and their optimal MSE; both are scalars.

• $y_{t|t-1}$ and $\sigma^2_{\epsilon_t}$ vary with time; in original model they were time invariant.

• If $f(y_1)$ is constant, prediction errors $=$ innovations in y_t, and $\sigma^2_{\epsilon_t} = \sigma^2_\epsilon,\ \forall t > 1$.

Example 10 $y_t = Ay_{t-1} + e_t, \quad |A| < 1,\ e_t \sim \ iid\ \mathbb{N}(0, \sigma^2_e)$. *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, $y_{t|t-1} \sim (Ay_{t-1}, \sigma^2_e)$. *Hence,* $\epsilon_t = y_t - y_{t|t-1} = y_t - Ay_{t-1} = e_t$ *and* $\sigma^2_e = \sigma^2_{\epsilon_t}$ *for* $t \geq 2$. *The unconditional of* y_1 *is* $y_1 \sim \mathbb{N}(0, \frac{\sigma^2_e}{1-A^2})$. *Setting* $\epsilon_1 = y_1$*:*

$$
\begin{aligned}
\mathcal{L}(y_1, \ldots, y_T|\phi) &= \sum_{t=2}^{T} \mathcal{L}(y_t|y_{t-1}, \ldots, y_1, \phi) + \mathcal{L}(y_1|\phi) \\
&= -\frac{T}{2}(\ln(2\pi) + \ln(\sigma_e^2)) - \frac{1}{2}\sum_{t=2}^{T} \frac{(y_t - Ay_{t-1})^2}{\sigma_e^2} \\
&+ \frac{1}{2}(\ln(1 - A^2) - \frac{(1 - A^2)y_1^2}{\sigma_e^2}) \qquad (21)
\end{aligned}
$$

Hence σ_{ϵ_t} *is a constant and* $t \geq 2$, *and* $\sigma_{\epsilon_1}^2 = \frac{\sigma_e^2}{1-A^2}$.

• Conditioning on initial observations eliminates nonlinearities. Conditional decomposition useful to estimate models with MA terms (typically difficult to deal with). As $T \to \infty$, contribution of the first observation to the likelihood negligible and exact and conditional ML coincide.

• If model has constant coefficients, the errors normally distributed and initial observations given, maximum likelihood and OLS estimators coincide (not if the model has MA terms).

• Multivariate decomposition (y_t is $m \times 1$).

$$\begin{aligned} \mathcal{L}(y_1, \ldots, y_t, \phi) &= -\frac{Tm}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{t} \ln |\Sigma_{t|t-1}| \\ &- \frac{1}{2} \sum_{t=1}^{T} (y_t - y_{t|t-1}) \Sigma_{t|t-1}^{-1} (y_t - y_{t|t-1}) \end{aligned} \tag{22}$$

where $\epsilon_t = y_t - y_{t|t-1} \sim \mathbb{N}(0, \Sigma_{t|t-1})$ and $y_1 \sim \mathbb{N}(\bar{y}_1, \Sigma_1)$; $\epsilon_1 = y_1 - \bar{y}_1$.

• The Kalman filter can be used to compute the likelihood function (it produces ϵ_t and $\Sigma_{t|t-1}$) for any model with a state space representation.

• Maximization/filtering (EM) algorithm.

Let $\phi = [vec(x_1), vec(x_2), vec(\mathbb{D}_1), vec(\mathbb{D}_0), vec(\mathbb{D}_2)]$

1) Choose ϕ^0. To choose the initial values of α: run an OLS regression on the constant coefficient version of the model (consistent for average), or on the sample $[\tau, 0]$.

2) Run KF for each t.

3) Save $\epsilon_t = y_t - y_{t|t-1}$ and $\Sigma_{t|t-1}$. Construct the likelihood (22) each t.
(For large scale models, use Choleski factor $\Sigma_{t|t-1} = \mathcal{P}_t \mathcal{P}_t'$).

4) Update ϕ^0 using any of the methods described in next section.

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

Once converged, standard errors of estimates are obtained from square root of the diagonal elements of Hessian $H(\phi_{ML})$.

- ϕ_{ML} consistent;

- $T^{0.5}(\phi_{ML} - \phi_0) \xrightarrow{D} \mathbb{N}(0, T^{-1}\mathcal{I}^{-1})$; $\mathcal{I} = -E(\sum_t \frac{\partial^2 \ln \mathcal{L}}{\partial \phi \partial \phi'}|_{\phi=\phi_0})$.

For this to occur we need:

i) the state equation defines a covariance stationary process. For constant coefficient models: sufficient that the roots of $\mathbb{D}_{1t} < 1$.

ii) The exogenous variables are covariance stationary, linearly regular.

iii) ϕ_0 is not on the boundary of the parameter space

iv) The likelihood is, roughly, quadratically.

• If distribution of errors misspecified: KF estimates still consistent. (Intuition: if T large, assuming normality not bad).

• Estimates of asymptotic covariance matrix:

i) $var_1(\phi) = -\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial \phi \partial \phi'}|_{\phi=\phi_{ML}}$.

ii) $var_2(\phi) = -\sum_t (\frac{\partial \ln \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi_{ML}})(\frac{\partial \ln \mathcal{L}(\phi)}{\partial \phi}|_{\phi=\phi_{ML}})'$.

iii) (QML) $var_3(\phi) = \frac{1}{T}((\frac{1}{T}var_1(\phi))(\frac{1}{T}var_2(\phi))^{-1}(\frac{1}{T}var_1(\phi)))^{-1}$.

• Hypothesis testing:

- t-tests

- LR tests, e.g. $-2(\mathcal{L}_U - \mathcal{L}_R) \sim \chi^2(\nu)$

- LM test: $\frac{1}{T}[\sum_t \frac{\partial \ln \mathcal{L}(\phi)}{\partial \phi}]'(\mathcal{I})^{-1}[\sum_t \frac{\partial \ln \mathcal{L}(\phi)}{\partial \phi}] \sim \chi^2(\nu)$, $\nu =$ number of restrictions.

Example 11 *For an ARMA(2,1) model, if* $D_1 = 0$*, conditional ML estimates of* $A = [A_1, A_2]'$ *solve* $Ax'x = x'y$*, where* $x_t = [y_{t-1}, y_{t-2}]$*,* $x = [x_1, \ldots x_t]'$*. If* $D_1 \neq 0$*, the equations are nonlinear and no closed form solution exists. Impose* $D_1 = 0$ *in estimation and test if restriction holds.*

5 Methods to maximize functions

Grid search

- Feasible when the dimension of ϕ is small.

- Advantage: no derivatives needed.

- Disadvantage: if function not globally concave, multiple peaks, may stop at local maximum.

Use as initial conditions for other algorithms.

Simplex method

- If $\ln \mathcal{L}(\phi_m) = \max_{j=1,m+1} \ln \mathcal{L}(\phi_j)$ replace ϕ_m by $\varrho\phi_m + (1-\varrho)\bar{\phi}$, where $\bar{\phi}$ is the centroid of (m+1) points.

- Advantage: fact, no derivatives needed, use when gradient methods fail.

- Disadvantage: no standard errors are available.

Gradient methods:

a) Steepest ascent: $\phi^l = \phi^{l-1} + \frac{1}{2\lambda} g(\phi^l)$ where $g(\phi^l) = \frac{\partial \ln \mathcal{L}(\phi=\phi^l)}{\partial \phi}$, λ is the Lagrangian multiplier of the problem

$\max_{\phi^l} \ln \mathcal{L}(\phi^l)$ subject to $(\phi^i - \phi^{l-1})'(\phi^l - \phi^{l-1}) = \kappa$.

If $\phi^l \approx \phi^{l-1}$, $g(\phi^l) = g(\phi^{l-1})$ then $\phi^l = \phi^{l-1} + \varrho g(\phi^{l-1})$, $\varrho \approx 10^{-5}$. Problem: it requires a lot of iterations.

b) Newton-Raphson: applicable if $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial \phi \partial \phi'}$ exists and $\ln \mathcal{L}(\phi)$ is concave (i.e. $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial \phi \partial \phi'}$ positive definite).

$$\begin{aligned} \ln \mathcal{L}(\phi) &= \ln \mathcal{L}(\phi^0) + g(\phi^0)(\phi - \phi^0) \\ &- 0.5(\phi - \phi^0)' \frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial \phi \partial \phi'}(\phi^0)(\phi - \phi^0) \end{aligned} \tag{23}$$

Maximizing $\ln \mathcal{L}(\phi)$ with respect to ϕ leads to:

$$\phi^l = \phi^{l-1} + (\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial \phi \partial \phi'}(\phi^l))^{-1} g(\phi^l) \tag{24}$$

If likelihood quadratic (24) generates MLE in one step. If close to quadratic $\rightarrow$ good properties. If far from quadratic worse than steepest ascent.

c) Hybrid: $\phi^l = \phi^{l-1} + \varrho(\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}(\phi^l))^{-1} g(\phi^l)$, $\varrho > 0$.

d) Modified Newton-Raphson: (b) requires inversion of $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$. Modified method uses $\frac{\partial g(\alpha)}{\partial\alpha} \approx \frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$. Let Σ^l be an estimate of $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ at iteration l.

$$(\Sigma^l) = (\Sigma^{l-1})^{-1} - \frac{(\Sigma^{l-1})^{-1}(\Delta g^l)(\Delta g^l)'(\Sigma^{l-1})^{-1}}{(\Delta g^l)'(\Sigma^{l-1})^{-1}(\Delta g^l)} + \frac{(\Delta\phi^l)(\Delta\phi^l)'}{(\Delta g^l)'(\Delta\phi^l)} \tag{25}$$

where $\Delta\phi^l = \phi^l - \phi^{l-1}$, $\Delta g(\phi^l) = g(\phi^l) - g(\phi^{l-1})$. If likelihood is quadratic and l large, $\lim_{l\to\infty} \phi^l = \phi_{ML}$ and $\lim_{l\to\infty} \Sigma^l = (\frac{\partial^2 \ln \mathcal{L}(\phi_{ML})}{\partial\phi\partial\phi'})^{-1}$. Standard error= diagonal elements of Σ^l evaluated at ϕ_{ML}.

e) Scoring Method. Uses the information matrix $E\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ in place of $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ where the expectation is evaluated at ϕ^{l-1}. Information matrix approximation convenient: simpler than Hessian.

f) Gauss-Newton method. Approximates $\frac{\partial^2 \ln \mathcal{L}(\phi)}{\partial\phi\partial\phi'}$ with a function of $(\frac{\partial e}{\partial\phi}|_{\phi^l})'(\frac{\partial e}{\partial\phi}|_{\phi^l})$, where ϕ_l is the value of ϕ at iteration l and e_t are the errors in the model. For constant state space models, the approximation is proportional to the vector of regressor constructed using the right hand side variables of both the state and the measurement equations. If the model is linear in parameters: Gauss-Newton$=$ Scoring.

Numerical Methods.

Work in situations where gradient methods fail. In particular, when the objective function is not smooth, not continuous, can have local maxima or large flat areas. Need only the function to be bounded (otherwise no maximum).

- Simulated Annealing. Procedure has two loops. Internal loop to explore the function you want to maximize. External loop to zoom-in in the area where the first loop has found local maxima to find the global one.

  Idea:

  1) Given a parameter vector, a new vector of candidates is generated with a Random Walk Metropolis algorithm using a uniform distribution on the jump and the value of the objective function

is found at the old and new parameter values. Given a T (a parameter to be chosen by the investigator) accept the move if $exp(-\Delta logL/T)$ is larger than a uniform random variable drawn from a (0,1) distribution, otherwise reject and make a new draw.

2) Repeat step 1, J times, starting from different initial conditions.

3) Let $T_{new} = \rho T, \rho \in (0, 1)$. Repeat steps 1)-2).

4) Repeat steps 1)-3) starting at the optimnal parameter values found in 1)-2). Since T is smaller, the probability to reject a draw is larger. Continue until $|logL - logL^*| < \epsilon$ where * indicates the maximum at the previous iteration.

Problems:

i) The algorithm does not work on non-convex sets so need to put an upper and lower bound on the generation of candidates in 1) and if a candidate goes outside the bounds, pick a random point in the interval.

ii) Time consuming.

iii) Lots of parameters needs to be set by investigator. Needs trials and errors (see Andreasan, 2010).

- Genetic algorithm. Tries to approximate the countours of the function you want to maximize numerically and get better and better approximation with iterations.

  Idea:

  1) Start from $\sigma^0 = 1, C^0 = I$.

  2) Generate M points in generation g+1: $x_i^{g+1} \sim N(x_w^g, (\sigma^g)^2 C^g)$, $i = 1, \ldots, M$.. Can put bounds on x^{g+1}. If draws are outside the bounds, resample until draws are inside. Compute the objective function at these points.

  3) Compute $x_w^{g+1} = \sum_{i=1}^{M} 1_{<M} \, w_i x_i^{g+1}$, i.e. use a weighted average of a subset of the points you have drawn. Update $(\sigma^g)^2$ and C^g using the previous estimate plus a piece which depends on the correlation among generations and a piece correcting for the dimensionality of x_i.

4) Repeat steps 2)-3). Continue until the value of the objective function at the new set of points from generation $g + 1$ is not different than the value in the previous generation (or particular average across previous generations). Here take the sup across dimensions.

Problems:

i) Many free parameters to be chosen (for some standard choices see Andreasan (2010).

ii) Could be very computationally intensive.

iii) Works also if the objective function is not convex but better to convexify the space by resampling.

6 ML estimation of DSGE models

• Log linearized solution of a DSGE model is

$$y_{2t} = \mathcal{A}_{22}(\theta) y_{2t-1} + \mathcal{A}_{21}(\theta) y_{3t} \quad (26)$$

$$y_{1t} = \mathcal{A}_{1}(\theta) y_{2t} = \mathcal{A}_{11}(\theta) y_{2t-1} + \mathcal{A}_{12}(\theta) y_{3t} \quad (27)$$

y_{2t} = states and the driving forces, y_{1t} = controls, y_{3t} shocks. $\mathcal{A}_{ij}(\theta)$, $i, j = 1, 2$ are time invariant (reduced form) matrices which depend on θ, the structural parameters of preferences, technologies, policies, etc. There are cross equation restrictions since $\theta_i, i = 1, \ldots, n$ appears in more than one entry of these matrices.

- (27) is a **singular** state space system.

Example 12 *(Sticky price model) Assume* $U(c_t, N_t, m_t) = \ln c_t + \ln(1 - N_t) + \frac{m_t^{1-\epsilon}}{1-\epsilon}$, *and the production function* $y_t = k_t^{\alpha} N_t^{1-\alpha} \zeta_t$. *Set* $N^{ss} = 0.33$, $\pi^{ss} = 1.005$, $\beta = 0.99$, $(\frac{C}{GDP})^{ss} = 0.7$, $\epsilon = 7$ *(consumption elasticity of money demand),* $\gamma_p = 0.75$ *(on average firms change prices every three quarters). Persistence of technology disturbances=0.95; persistence of monetary policy shocks=0.75 Parameters of policy rule:* $\varpi_1 = 0.5; \varpi_2 = 1.6$. *Then the decision rules are:*

$$\begin{bmatrix} \widehat{\pi}_t \\ \widehat{k}_t \\ \widehat{c}_t \\ \widehat{y}_t \\ \widehat{N}_t \\ \widehat{w}_t \\ \widehat{i}_t \\ \widehat{m}_t \end{bmatrix} = \begin{bmatrix} 0.12 & 0.02 \\ 1.36 & 0.90 \\ 0.80 & 0.53 \\ 16.95 & -0.51 \\ 26.49 & -1.37 \\ 0.80 & 0.53 \\ 10.07 & -0.25 \\ 1.30 & -0.11 \end{bmatrix} \begin{bmatrix} \widehat{\pi}_{t-1} \\ \widehat{k}_{t-1} \end{bmatrix} + \begin{bmatrix} -0.03 & 0.01 \\ 0.20 & -0.01 \\ 0.44 & -0.01 \\ 2.73 & -0.19 \\ 2.70 & -0.31 \\ 0.44 & -0.01 \\ 1.36 & 0.90 \\ 0.12 & 0.12 \end{bmatrix} \begin{bmatrix} \widehat{\epsilon}_{1t} \\ \widehat{\epsilon}_{3t} \end{bmatrix}$$

$\widehat{\epsilon}_{1t}$ = *technological disturbance;* $\widehat{\epsilon}_{3t}$ = *monetary disturbance.*

• (26)-(27) is general; certainty equivalence not required, needs not be the solution to the model; can be used also in partial equilibrium models (Watson (1989))

Example 13 $E_t y_{t+1} = \alpha y_t + x_t$ *where* $x_t = \rho x_{t-1} + e_t^x$, x_0 *given. (e.g. in a New-Keynesian Phillips curve:* $x_t = mc_t$, *valuation equation:* $x_t = sd_t$*). Note that* $x_t = E_{t-1}x_t + e_t^x,\ y_t = E_{t-1}y_t + e_t^y$ *where* $E_t x_{t+1} = \rho x_t = \rho(E_{t-1}x_t + e_t^x)$ *and* $E_t y_{t+1} = \alpha(E_{t-1}y_t + e_t^y) + (E_{t-1}x_t + e_t^x)$.
To fit the model into (27) set $y_{1t} = [x_t, y_t],\ y_{2t} = [E_{t-1}x_t, E_{t-1}y_t],\ y_{3t} = [e_t^x, v_t]$, *where* $v_t = e_t^y - E(e_t^y|e_t^x) = e_t^y - \kappa e_t^x$, $\mathcal{A}_{11}(\theta) = I$, $\mathcal{A}_{12}(\theta) = \begin{bmatrix} 1 & 0 \\ \kappa & 1 \end{bmatrix}, \mathcal{A}_{22}(\theta) = \begin{bmatrix} \rho & 0 \\ 1 & \alpha \end{bmatrix}, \mathcal{A}_{21}(\theta) = \begin{bmatrix} \rho & 0 \\ 1+\alpha\kappa & \alpha \end{bmatrix}$. *Here* $\theta = (\alpha, \rho, \kappa, \sigma_e^2, \sigma_v^2)$.

- DSGE-ML algorithm:

  i) Pick $\theta = \theta^0, \sigma_{y_3} = \sigma_{y_3}^0, y_{20}$.

  ii) Solve the model and run the Kalman filter.

  iii) Compute the likelihood and maximize it with respect to θ, σ_{y_3}.

  iv) Repeat i)-iii) until convergence. Read standard errors off the Hessian.

Difficult issues:

• Parameters need to be identifiable. Difficult to say if all are, and if they are not, which are identifiable if $dim(\theta)$ large - some parameters may enter only the steady states (not taken into account in the likelihood) or in combination with others (Iskrev (2008)).

In practice, arbitrarily choose one set (θ_1) and estimate others (θ_2). Problem of consistency and asymptotic distribution of θ_2. Better: jointly estimate θ_1, θ_2 using scores and moments (add conditions).

• (27) is the solution to a DSGE model. Since solution mixes up the information content of all equations, all parts of the model must true for ML estimation to be consistent. Can we assume that?

• Unconstrained maximization often leads to estimates on boundary. Transform parameter space so that support is the real line or do constrain maximization (see later on).

• Singularity problem: shock vector smaller than endogenous variables vector.

Solution 1: add serially uncorrelated, contemporaneously correlated measurement errors to some variables (see e.g. Altug (1989), Kim (2000)). Ireland (2003): VAR(1) measurement error.

Solution 2: drop variables. Which ones? Need to make sure that kept ones have information (they are ancillary to the parameters). Same problem when there are too many potential instruments - IV estimates may be different, all of them inefficient.

Two ways of doing this: a) write likelihood of m equations (hopefully some variables are non-observables); b) reduce solution to a m-equations model.

In (27), if both y_{2t} and y_{1t} are used, solution is also a restricted VAR(1)). If reduced to only y_{1t}, how the solution looks like?

i) If $\mathcal{A}_{12}$ is invertible

$$y_{3t} = \mathcal{A}_{12}^{-1}(y_{1t} - \mathcal{A}_{11}y_{2t-1})$$

$$y_{2t} = \mathcal{A}_{22}y_{2t-1} + \mathcal{A}_{21}\mathcal{A}_{12}^{-1}(y_{1t} - \mathcal{A}_{11}y_{2t-1})$$

$$(1 - (\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{12}^{-1}\mathcal{A}_{11})L)y_{2t} = \mathcal{A}_{21}\mathcal{A}_{12}^{-1}y_{1t}$$

If $\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{12}^{-1}\mathcal{A}_{11}$ has all eigenvalues less than 1

$$y_{2t} = (1 - (\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{12}^{-1}\mathcal{A}_{11})L)^{-1}\mathcal{A}_{21}\mathcal{A}_{12}^{-1}y_{1t}$$

and

$$y_{1t} = \mathcal{A}_{11}(1 - (\mathcal{A}_{22} - \mathcal{A}_{21}\mathcal{A}_{12}^{-1}\mathcal{A}_{11})L)^{-1}\mathcal{A}_{21}\mathcal{A}_{12}^{-1}y_{1t-1} + \mathcal{A}_{12}y_{3t} \quad (28)$$

(26)-(27) produce a VAR(∞) for y_{1t}.

ii) If $\mathcal{A}_{11}$ is invertible

$$y_{2t-1} = \mathcal{A}_{11}^{-1}(y_{1t} - \mathcal{A}_{12}y_{3t})$$

$$\mathcal{A}_{11}^{-1}(y_{1t+1} - \mathcal{A}_{12}y_{3t+1}) = \mathcal{A}_{22}\mathcal{A}_{11}^{-1}(y_{1t} - \mathcal{A}_{12}y_{3t}) + \mathcal{A}_{21}y_{3t}$$

$$y_{1t+1} = \mathcal{A}_{11}\mathcal{A}_{22}\mathcal{A}_{11}^{-1}y_{1t} + (\mathcal{A}_{11}\mathcal{A}_{21} - \mathcal{A}_{11}\mathcal{A}_{22}\mathcal{A}_{11}^{-1}\mathcal{A}_{12})y_{3t} + \mathcal{A}_{12}y_{3t+1})$$

$$y_{1t+1} = \mathcal{A}_{11}\mathcal{A}_{22}\mathcal{A}_{11}^{-1}y_{1t} + (I + (\mathcal{A}_{11}\mathcal{A}_{21}\mathcal{A}_{12}^{-1} - \mathcal{A}_{11}\mathcal{A}_{22}\mathcal{A}_{11}^{-1})L)y_{4t+1}$$

where $y_{4t} \sim (0, \mathcal{A}_{12}'\mathcal{A}_{12})$

(26)-(27) produce a VARMA(1,1) for y_{1t}.

iii) Final form computations. From ii) $y_{1t} = \mu_1 y_{1t-1} + u_t + \nu_1 u_{t-1}$.

For any two elements y_{1t}^1, y_{1t}^2

$$\begin{bmatrix} 1-\mu_{11}L & -\mu_{12}L \\ -\mu_{21}L & 1-\mu_{22}L \end{bmatrix} \begin{bmatrix} y_{1t}^1 \\ y_{1t}^2 \end{bmatrix} = \begin{bmatrix} 1-\nu_{11}L & \nu_{12}L \\ \nu_{21}L & 1+\nu_{22}L \end{bmatrix} \begin{bmatrix} u_t^1 \\ u_t^2 \end{bmatrix};$$

$$det(\nu(L)) = (1+\nu_{11}L)(1+\nu_{22}L) - \nu_{12}\nu_{21}L^2$$

$$\begin{bmatrix} 1+\nu_{22}L & -\nu_{21}L \\ -\nu_{12}L & 1+\nu_{11}L \end{bmatrix} \begin{bmatrix} 1-\mu_{11}L & -\mu_{12}L \\ -\mu_{21}L & 1-\mu_{22}L \end{bmatrix} \begin{bmatrix} y_{1t}^1 \\ y_{1t}^2 \end{bmatrix} = det(\nu(L)) \begin{bmatrix} u_t^1 \\ u_t^2 \end{bmatrix}$$

(26)-(27) produce a VARMA(2,2) for (y_{1t}^1, y_{1t}^2).

If rank of $\nu(L)$ is reduced, write

$$\nu(L) = \begin{bmatrix} 1 + \nu_{11}L & \alpha\nu_{11}L \\ \nu_{21}L & 1 + \alpha\nu_{21}L \end{bmatrix} \text{ some } \alpha.$$

Then $det(\nu(L)) = 1 + (\nu_{11} + \nu_{21})L$ and (26)-(27) produces VARMA(2,1) for (y_{1t}^1, y_{1t}^2).

- Model Validation (i), Statistical tests:

  (a) t-test, stability tests, Likelihood ratio test (restricted VAR (model) vs unrestricted VAR (data)), forecasting tests (see Ireland (2001, 2004)).

  (b) Cross equation restrictions.

Example 14 *Hybrid Philips curve (Gali and Gertler (1999))*

$$\pi_t = \alpha_1 E_t \pi_{t+1} + \alpha_2 \pi_{t-1} + \alpha_3 mc_t + e_t$$

e_t *measurement error. Let* $\mathbb{Y}_t = \mathbb{A}\mathbb{Y}_{t-1} + E_t$ *be the companion form where* $\mathbb{Y}_t$ *be 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, model implies* mq *restrictions of the form*

$$\mathcal{S}_2[\mathbb{A} - \alpha_1 \mathbb{A}^2 - \alpha_2 I] = \alpha_3 \mathcal{S}_2 \mathbb{A}$$

If $q = 1$, $\mathbb{Y}_t = (\pi_t$, *labor share*$_t)$ *and* A_{ij} *are VAR parameters:*

$$\begin{aligned} A_{12} - \alpha_1 A_{12} A_{11} - \alpha_1 A_{22} A_{12} &= \alpha_3 A_{11} \\ A_{22} - \alpha_1 A_{21} A_{12} - \alpha_1 A_{22}^2 - \alpha_2 &= \alpha_3 A_{21} \end{aligned} \tag{29}$$

Four unknown and two equations: solve for, e.g., A_{21} *and* A_{22} *as a function of* A_{11} *and* A_{12}.

Idea:(29) requires that expectations of real marginal costs and inflation given by a VAR to be consistent with the dynamics of the model. To impose (29) express VAR coefficients of the inflation equation as a function of the remaining $(m-1)mq$ *VAR coefficients and the parameters of the theory.*

Restricted likelihood is $\mathcal{L} \propto -\frac{T}{2}[\ln|\Sigma_\epsilon| + m$ *where* $\Sigma_\epsilon = \frac{1}{T}\sum_t \hat{\epsilon}_t(\theta)\hat{\epsilon}_t(\theta)'$ *is the variance covariance matrix of the VAR errors,* $\epsilon_t = y_t - \sum_t A_i(\theta)y_{t-i}$ *and where the constrained parameters necessary to compute* $\hat{\epsilon}_t$ *are obtained for each* t *from (29).*

To tests restrictions: compare likelihood of restricted (VAR) model and unrestricted VAR.

Cross equation restriction tests are limited information tests:

- other relationships disregarded (e.g. Euler equation).

- depend on VAR representation for the data.

- expectations must be the same as linear projections.

• Model Validation (ii); Economic tests. Given θ_{ML} :

(a) compute conditional moments (impulse responses/variance decompositions/historical decompositions) and standard errors.

(b) compute unconditional moments (variability, cross correlations) and standard errors.

(c) compute welfare measures, costs due to stochastic policy, etc. and standard errors.

7 Frequency domain maximum likelihood

- Kalman filter is convenient but there are other methods to compute the likelihood which are equally convenient.

- For DSGE models one may want to estimate the parameters at business cycle frequencies (rather than all frequencies).

- A way to do so is to express the likelihood of the model in frequency domain. The frequnecy domain approximation to the likelihood can be combined with a prior to yield a posterior distribution in a standard Bayesian analysis.

Recall: the (log-) linearized solution of a DSGE model is

$$
\begin{aligned}
y_{2t} &= \mathcal{A}_{22}(\theta) y_{2t-1} + \mathcal{A}_{21}(\theta) y_{3t} && (30)\\
y_{1t} &= \mathcal{A}_{1}(\theta) y_{2t} = \mathcal{A}_{11}(\theta) y_{2t-1} + \mathcal{A}_{12}(\theta) y_{3t} && (31)
\end{aligned}
$$

y_{2t} are states and the driving forces, y_{1t} are observable controls, y_{3t} shocks. $\mathcal{A}_{ij}(\theta)$,
$i, j = 1, 2$ are time invariant (reduced form) matrices which depend on θ, the structural parameters.

The log-likelihood of the observable y_{1t}'s is

$$
\begin{aligned}
L(\theta|y_t) &= -(T/2)(\ln(2\pi) - \ln|\Sigma_{y_3}|) \\
&- \frac{1}{2}(y_{1t} - \mathcal{A}_1(\theta)(1 - \mathcal{A}_{22}(\theta)\ell)^{-1}\mathcal{A}_{21}(\theta)y_{3t})'(\Sigma_{y_3})^{-1} \\
&\times (y_{2t} - \mathcal{A}_1(\theta)(1 - \mathcal{A}_{22}(\theta)\ell)^{-1}\mathcal{A}_{21}(\theta)y_{1t}) && (32)
\end{aligned}
$$

The spectral density of y_{1t} is

$$\begin{aligned} G_{y_1,\theta}(\omega_j) &= \frac{1}{2\pi}\mathcal{A}_1(\theta)(1-\mathcal{A}_{22}(\theta)e^{-i\omega_j})^{-1}\mathcal{A}_{21}(\theta)\Sigma_{y_3} \\ &\times \mathcal{A}_{21}(\theta)'(1-\mathcal{A}_{22}(\theta)e^{'i\omega_j})^{-1}\mathcal{A}_1(\theta)' \end{aligned} \tag{33}$$

Following Sargent and Hansen (1993) we can approximate using log-likelihood using the spectral density in the the following way:

$$L(\theta|y_{1t}) = A_1(\theta) + A_2(\theta) \tag{34}$$

$$A_1(\theta) = \frac{1}{\pi}\sum_{\omega_j} \log\det G_{y_1,\theta}(\omega_j) \tag{35}$$

$$A_2(\theta) = \frac{1}{\pi}\sum_{\omega_j>\omega_0} \text{trace}\,[Gy_{1,\theta}(\omega_j)]^{-1}F(\omega_j) \tag{36}$$

where $\omega_j = \frac{2\pi j}{T}, j = 0, 1, \ldots, T-1$, are Fourier frequencies, $Gy_{1,\theta}(\omega_j)$ is defined above and $F(\omega_j)$ is the data based spectral density of y_{1t}.

- The first term is the one-step ahead forecast error matrix (sum of the prediction errors across frequencies);

- The second term is a penalty function. It measures deviations of the model-based spectral density from the data-based spectral density at various frequencies (frequency ω_0 is excluded).

Since $F(\omega_j)$ is not available we estimate it with the periodogram $I(\omega_j)$ which is a consistent estimator (as T grows). The periodogram is computed as

$$I(\omega_j) = \frac{1}{T} q(\omega_j) q(\omega_j)' \tag{37}$$

where $q(\omega_j) = \sum_t y_{1t} e^{-i\omega_j t}$

Note that in case the steady states are included in the representation, i.e. the model is for the level of y_{1t} there is a third term to the likelikood approximation which is given by:

$$A_3(\theta) = (E(y) - \mu(\theta))Gy_{1,\theta}(\omega_0)^{-1}(E(y) - \mu(\theta)) \tag{38}$$

where $\mu(\theta)$ the model based mean of y_{t1} and $E(y_1)$ the unconditional mean of the data. This term is another penalty function, weighting deviations of model-based from data-based means, with the spectral density matrix of the model at frequency zero.

The elements in $A_1(\theta)$ and $A_2(\theta)$ are asymptotically uncorrelated. Thus, one can include only the elements associated to the frequencies of interest

- Can also estimate the model using different frequencies and check how the choice impact on parameter estimates. In this case $A_1(\theta)$ and $A_2(\theta)$ become

$$A_1(\theta)^{\dagger} = \frac{1}{\pi}\sum_{\omega_j} w(\omega_j) \log\det G_{y_1,\theta}(\omega_j) \tag{39}$$

$$A_2(\theta)^{\dagger} = \frac{1}{\pi}\sum_{\omega_j > \omega_0} w(\omega_j)\text{trace}[Gy_{1,\theta}(\omega_j)]^{-1}F(\omega_j) \tag{40}$$

where $w(\omega_j)$ is an indicator function, equal to 1 for the frequencies included and equal to 0 for the frequencies which are excluded.

Example 15 *Christiano-Viggfusson, JME, 2003. Estimate a number of RBC models with frequency domain ML.*

Table 1: Weighted Likelihood Estimation Results, RBC Model

Frequencies	θ	δ	σ_z	λ	λ_w	Observations
						Used
High	0.25	0.99	0.0126	9.6	3.7	50%
Business Cycle	0.51	0.99	0.0170	26.1	2.3	43%
Low	0.15	0	0.0100	37.3	-0.2	7%
All	0.37	0.73	0.0144	8.5	8.5	100%

- **Notes**: These are the results of estimating the unrestricted RBC model by weighted maximum likelihood (i.e., by maximizing (2.3)). Low frequencies: $v_j = 1$ only for v_j's that belong to frequencies corresponding to periods 8 years and up. Business cycle frequencies: $v_j = 1$ only for v_j's that belong to frequencies corresponding to periods 1 to 8 years. High frequencies: $v_j = 1$ only for v_j's that belong to frequencies corresponding to periods 2 quarters to 1 year; All frequencies: $v_j = 1$ for all j. Percent of observations used: fraction of $j \in \{0, 1, ..., T-1\}$ equal to unity in the weighted likelihood estimation. λ: likelihood ratio statistic based on all frequencies. λ_w: likelihood ratio statistic based only on the indicated subinterval of frequencies.

8 Examples

8.1 Example 1: Gali (1999, AER)

$$\begin{aligned}
\Delta p_t &= e_{3t-1} - (1-a_m)e_{1t-1} && (41)\\
\Delta gdp_t &= \Delta e_{3t} + a_m e_{1t} - (1-a_m)e_{1t-1} && (42)\\
n_t &= \frac{1}{\wp}e_{3t} - \frac{1-a_m}{\wp}e_{1t} && (43)\\
\Delta np_t &= (1-\frac{1}{\wp})\Delta e_{3t} + (\frac{1-a_m}{\wp} + a_m)e_{1t} + (1-a_m)(1-\frac{1}{\wp})e_{1t-1} && (44)
\end{aligned}$$

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

e_{1t}= technology shock, e_{3t}= monetary shock, a_m = response to money to technology, η_1 = exponent of effort and hours in intermediate production function, η_2 = weight on hours in Cobb-Douglas, φ_i exponent on effort and hours in utility.

• two shocks ($\epsilon_{1t}, \epsilon_{3t}$); four variables ($\Delta p_t, \Delta y_t, \Delta np_t, n_t$);

• 11 free parameters ($\eta_1, \eta_2, \varphi_n, \varphi_{ef}, \beta, \sigma_1^2, \sigma_2^2$, $a_m, \vartheta_M, \vartheta_n, \vartheta_U$). Only a_m and $\wp$ identifiable together with σ_1^2 and σ_2^2.

• State space representation: $\alpha = [\epsilon_{1t}, \epsilon_{1t-1}, \epsilon_{3t}, \epsilon_{3t-1}, v_{1t}, v_{2t}]$, ($v_{1t}$ and v_{2t} measurement errors); $\Sigma_{v_1} = 0$,

$$x_{1t} = \begin{bmatrix} 0 & a_M - 1 & 0 & 1 & 0 & 0 \\ a_M & 1 - a_M & 1 & -1 & 1 & 0 \\ \frac{a_M - 1}{\wp} & 0 & \frac{1}{\wp} & 0 & 0 & 0 \\ \frac{1-a_M}{\wp} + a_M & \frac{(1-a_M)(\wp-1)}{\wp} & \frac{\wp-1}{\wp} & -\frac{\wp-1}{\wp} & 0 & 1 \end{bmatrix},$$

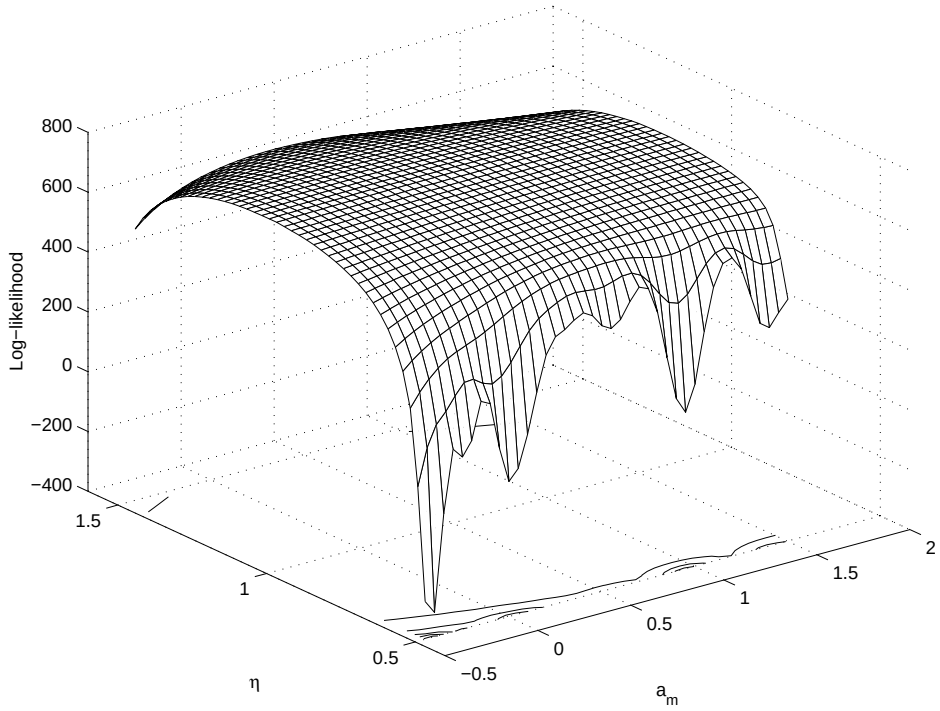
$$\mathbb{D}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbb{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

ML estimates, Canada 1980-2002

Data set	a_m	$\wp$	σ_1^2	σ_2^2			likelihood
$(\Delta np_t, n_t)$	0.5533	0.9998	1.06e-4	6.69e-4			704.00
$(\Delta y_t, n_t)$	-7.7336	0.7440	6.22e-6	1.05e-4			752.16
$(\Delta y_t, \Delta p_t)$	3.2007		1.26e-5	1.57e-4			847.12
	a_m	$\wp$	σ_1^2	σ_2^2	σ_{v1}^2	σ_{v2}^2	likelihood
$(n_t, \Delta np_t, \Delta y_t, \Delta p_t)$	-0.9041	1.2423	5.82e-6	4.82e-6	0.0236	0.0072	1336

p-values	$a_m = 0$	$\wp = 1$	$\wp = 1, a_m = -1.0$	$\wp = 1.2$
$(\Delta np_t, n_t)$	0.03	0.97	0.01	0.00
$(\Delta y_t, n_t)$	0.00	0.00	0.00	0.00
$(n_t, \Delta np_t, \Delta y_t, \Delta p_t)$	0.00	0.001	0.00	0.87

	$a_m = 0$	$a_m = 1$	$a_m = -1.0$
$(\Delta y_t, \Delta p_t)$	0.00	0.00	0.00



Likelihood surface

Cross covariances

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

8.2 Example 2: Is Government expenditure procyclical?

Agents max $E_t \sum_t \beta^t (\log C_t - \gamma N_t)$ subject to

$$
\begin{aligned}
G_t + C_t + I_t &= A_t k_t^{\eta} (\theta_t N_t)^{1-\eta} = Y_t && (45)\\
k_{t+1} &= (1-\delta) k_t + I_t && (46)\\
\log A_t &= (1-\rho_A) \log A + \rho_A \log A_{t-1} + e_{1t} && (47)\\
\log G_t &= (1-\rho_G) \log G + \rho_G \log G_{t-1} + \zeta Y_{t-1} + e_{2t} && (48)
\end{aligned}
$$

Parameters:$\beta, \gamma, \theta, \eta, \delta, A, \rho_A, \sigma_A, G, \rho_G, \sigma_G, \zeta$.

Interest: sign and magnitude of ζ

Use linearly detrended US data, 1948-2002 for (C, H, Y, I), add two measurement errors.

Parameter	Estimate	St. Err.
β	0.99	NA
γ	3.196	0.011
η	0.098	0.0001
θ	1.026	NA
δ	0.045	0.034
A	3.001	72.77
ρ_A	0.994	0.127
σ_A	32.02	0.021
G	1.047	0.024
ρ_G	0.685	0.001
σ_G	28.56	0.657
ζ	-2.012	0.032
σ_{1m}	54.85	0.827
σ_{2m}	62.56	0.992

Detrended data is not stationary!!!

8.3 Example 3: Does money matter for business cycles?

- Use a basic New-Keynesian model without capital.

- Allow

i) external habits in consumption: $x_t = c_t - hC_{t-1}$,

ii) real balances and consumption non-separable in utility: $U(x, \frac{M}{p}) - V(n)$;

iii) the growth rate of nominal balances enters the nominal interest rate determination: $R_t = f(y_{t-p}, \pi_{t-p}, \Delta M_{t-p})p \geq 0$.

The log-linearized conditions

$$
\begin{aligned}
\hat{y}_t &= \frac{1}{1+h}E_t\hat{y}_{t+1} + \frac{h}{1+h}\hat{y}_{t-1} - \frac{\omega_1}{1+h}((\hat{R}_t - E_t\hat{\pi}_{t+1}) - (\hat{a}_t - E_t\hat{a}_{t+1})) \\
&+ \frac{\omega_2}{1+h}((\mathbf{\hat{m}_t} - \mathbf{\hat{e}_t}) - (\mathbf{E_t\hat{m}_{t+1}} - \mathbf{E_t\hat{e}_{t+1}})) \qquad (49) \\
\hat{m}_t &= \gamma_1(\hat{y}_t - h\hat{y}_{t-1}) - \gamma_2\hat{R}_t + (1 - (R^s - 1)\gamma_2)\hat{e}_t \qquad (50) \\
\hat{\pi}_t &= \beta E_t\hat{\pi}_{t+1} + \psi(\frac{1}{\omega_1}(\hat{y}_t - h\hat{y}_{t-1}) - \hat{z}_t - \frac{\omega_2}{\omega_1}(\mathbf{\hat{m}_t} - \mathbf{\hat{e}_t})) \qquad (51) \\
\hat{R}_t &= \rho_r\hat{R}_{t-1} + (1-\rho_r)\rho_y\hat{y}_{t-p} + (1-\rho_r)\rho_\pi\hat{\pi}_{t-p} \\
&+ (\mathbf{1 - \rho_r})\boldsymbol{\rho_m\Delta}(\mathbf{\hat{m}_{t-p}} + \mathbf{\hat{\pi}_{t-p}}) + \hat{\epsilon}_t \qquad (52)
\end{aligned}
$$

where

$$\omega_1 = -\frac{U_1(x_t, \frac{m_t}{e_t})}{y^s U_{11}(x^s, \frac{m^s}{e^s})} \tag{53}$$

$$\omega_2 = -\frac{m^s}{e^s}\frac{U_{12}(x^s, \frac{m^s}{e^s})}{y^s U_{11}(x^s, \frac{m^s}{e^s})} \tag{54}$$

$$\gamma_2 = \frac{R^s}{(R^s-1)(m^s/e^s)}\left(\frac{U_2(x^s, \frac{m^s}{e^s})}{(R^s-1)e^s U_{12}(x^s, \frac{m^s}{e^s}) - R^s U_{22}(x^s, \frac{m^s}{e^s})}\right) \tag{55}$$

$$\gamma_1 = (R^s - 1 + R^s\omega_2\frac{y^s}{m^s})(\frac{\gamma_2}{\omega_1}) \tag{56}$$

$$\psi = \frac{\theta - 1}{\phi} \tag{57}$$

Data

- 1959:1 to 2008:2 for the US (FRED Data base)

- 1970:1 to 2007:4 for the Euro area (ECB)

- 1965:1 to 2008:2 for the UK (Bank of England)

- 1980:1 to 2007:4 for Japan (IMF and OECD data bases)

- Inflation = GDP deflator; Money = M2 (in UK M4), to be consistent with literature; Interest rates = 3 months rate.

Full sample estimates

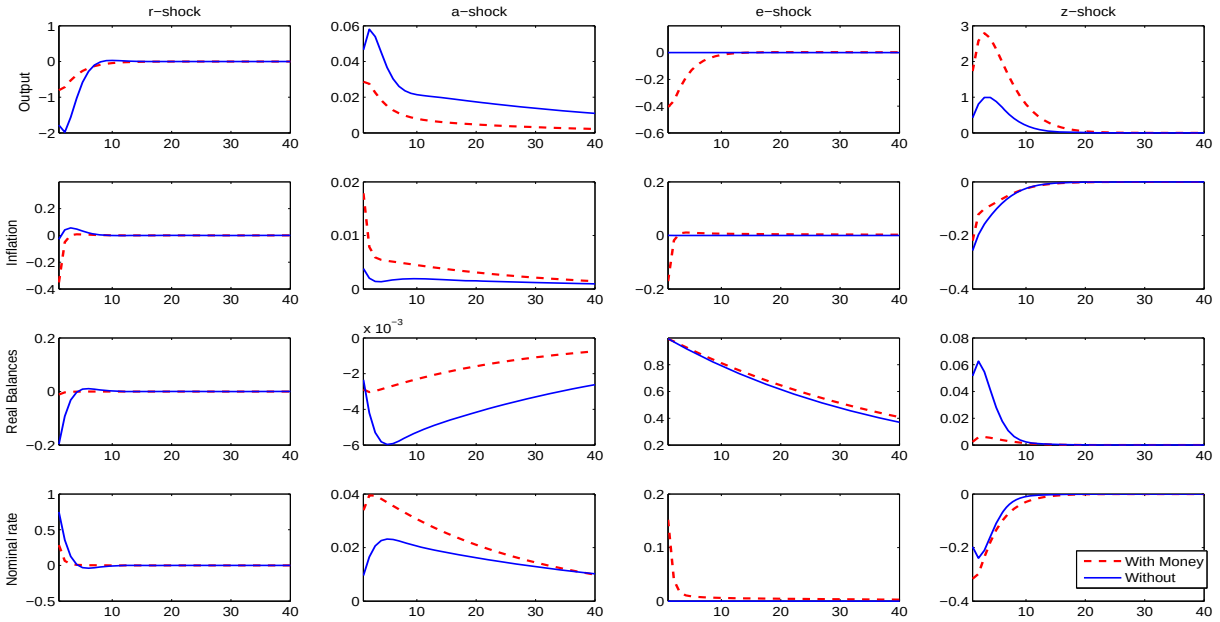
Parameter	US	Japan	EU	UK
ω_2	-0.511 (0.482)	-0.225 (0.135)	-0.290 (0.058)	0.174 (0.034)
ρ_m	1.578 (0.195)	1.047 (0.447)	1.071 (0.267)	-0.365 (0.475)
p	1	0	0	0
LR test p-value Log Bayes factor	0.00 17.53	0.00 22.95	0.00 26.56	0.00 42.72

- ω_2 :direct role of money. ρ_m the long run indirect effects of money.

- $p = 0$ contemporaneous rule, $p = 1$ lagged rule.

- Standard errors in parenthesis. LR test and Log Bayes factor test jointly $\omega_2 = 0, \rho_m = 0$. The LR test uses $2(\log L_u - \log L_r)$, the log Bayes factor is approximated by $(\log L_u - \log L_r) - 0.5(k_u - k_r) * \log(T)$, where k_j is the number of parameters in $j = u, r$. Log Bayes factor strongly significant if value > 10, weakly significant if value [2,10].

What is the economic relevance of money?



Responses to shocks US, Full sample

One step ahead historical decomposition of EU inflation

- Most of the fall is predictable in both cases.

- Without money: technology shocks much less important and monetary policy shocks relatively more important.

Tips

- Likelihood of DSGE models badly behaved. Start optimization many times from different initial conditions. Map the shape of the likelihood function to find the maximum.

- Use a "good" optimizer (e.g. csminwell.m is good, fminunc.m is bad).

- Explore well flat regions: there may be a spike somewhere.

- Check model misspecification. Likelihood bad if model is poorly specified in some dimensions.

- Small samples cause the likelihood to be flat.

9 Exercises

1) Suppose $y_t = x_t\alpha_t + v_{1t}$ and $\alpha_t = \alpha_1$ if $t < T_0$ and $\alpha_t = \alpha_2 > \alpha_1$ if $t \geq T_0$. Show what is the Kalman filter estimate of α. Is the Kalman filter optimal here? Why?

2) Suppose $y_t = e_t + \theta e_{t-1}$. Write down the prediction error decomposition for this model. Can I find θ treating y_1 as given? Why? Why not?

3) Can I estimate the parameters θ of a DSGE model using the following two step approach? Estimate $\mathcal{A}_{ij}$ with the Kalman filter from the data; find θ to minimize $||\widehat{\mathcal{A}}_{ij} - \mathcal{A}_{ij}(\theta)||$. Why? How does this compare to maximum likelihood estimate?