Estimating DSGE models via filtering to recover states: Part 1

Tom Holden

http://www.tholden.org/

PhD Macroeconomics, Semester 2

Outline of today's talk

- Motivation.
- The Kalman Filter.
- Kalman Smoothing.
- The Extended Kalman Filter.
- Introduction to non-linear filtering.
- The Kollman (2013) approach.

- Next week:
  - Numerical integration.
  - Numerical integration based filters (unscented, quadrature, particle).
  - Numerical integration based Bayesian estimation.

Reading for today

- Canova: “Methods for applied macroeconomic research”.
  - Chapter 6 covers the Kalman Filter, the prediction error decomposition, and various aspects of Maximum Likelihood Estimation.

Motivation (1/3)

- Suppose we observe $x_1, \ldots, x_T$, where:
$$x_t = \rho x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{NIID}(0, \sigma^2), \qquad x_0 = m_0.$$
  - where m_0 is a known parameter representing the point at which the process was started.
- This may be estimated by taking the joint maximum likelihood over ρ and σ.
- The log-likelihood takes the form:
$$-\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2 .$$
- Note that if m_0 was in fact unknown, we could treat it as a parameter, and its FOC would imply $\frac{\hat{\rho}}{\hat{\sigma}^2}(x_1 - \hat{\rho}\hat{m}_0) = 0$, i.e. $\hat{\varepsilon}_1 = 0$, and we are left with estimates that are asymptotically equivalent to the results of an OLS regression of $x_2, \ldots, x_T$ on $x_1, \ldots, x_{T-1}$. (Exercise: prove this.)

- However, in many situations it is neither case that we know the initial value of the process, nor that we are completely uncertain about it.
  - Instead, we might have non-degenerate prior beliefs about x_0.

Motivation (2/3)

- Suppose $x_0 \sim \mathrm{NID}(m_0, s_0^2)$.
  - E.g. suppose the process was actually started a long time before time 0, and so x_0 is a draw from the stationary distribution of x_t. In this case $m_0 = 0$ and $s_0^2 = \frac{\sigma^2}{1-\rho^2}$.
- Suppose for some t, $x_t \sim \mathrm{N}(m_t, s_t^2)$.
- Then, $x_{t+1} = \rho x_t + \varepsilon_{t+1} \sim \mathrm{N}(\rho m_t, \rho^2 s_t^2 + \sigma^2)$ and so $x_{t+1} \sim \mathrm{N}(m_{t+1}, s_{t+1}^2)$, with $m_{t+1} = \rho m_t$ and $s_{t+1}^2 = \rho^2 s_t^2 + \sigma^2$.
- Iterating this back gives: $x_t \sim \mathrm{N}\left(\rho^t m_0, \rho^{2t} s_0^2 + \frac{\sigma^2(1-\rho^{2t})}{1-\rho^2}\right)$, so x_t is both conditionally and unconditionally normally distributed.
- The log-likelihood in this case is:

$$-\frac{T}{2}\log 2\pi - \frac{1}{2}\log(\rho^2 s_0^2 + \sigma^2) - \frac{T-1}{2}\log\sigma^2 - \frac{(x_1 - \rho m_0)^2}{2(\rho^2 s_0^2 + \sigma^2)} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(x_t - \rho x_{t-1})^2 .$$

- With $s_0 = 0$ we get back the perfect information case from the previous slide, and in the limit as $s_0 \to \infty$ we are left with estimates that are identical to the results of an OLS regression of $x_2, \ldots, x_T$ on $x_1, \ldots, x_{T-1}$. (Exercise: prove this too.) For intermediate values though, $\hat{x}_1 \neq \hat{\rho} m_0$.

Motivation (3/3)

- Now suppose that rather than observing x_t, we instead observed $y_t = x_t + \eta_t$ where $\eta_t \sim \text{NIID}(0, \omega^2)$.
- Clearly y_t is still conditionally and unconditionally normally distributed. In fact, note:

$$y_t - \rho y_{t-1} = x_t - \rho x_{t-1} + \eta_t - \rho \eta_{t-1} = \varepsilon_t + \eta_t - \rho \eta_{t-1}.$$

  - So if we define $Z = [y_T - \rho y_{T-1} \quad \cdots \quad y_2 - \rho y_1]'$, and $\Sigma \coloneqq \mathbb{E}ZZ'$, then Σ is a tri-diagonal matrix with diagonal $\sigma^2 + (1-\rho^2)\omega^2$ and off-diagonals $-\rho\omega^2$.
- Now let L be the Cholesky decomposition of Σ, so $LL' = \Sigma$.
  - Then $\mathbb{E}(L^{-1}Z)(L^{-1}Z)' = L^{-1}\Sigma(L')^{-1} = L^{-1}LL'(L')^{-1} = I$.
  - So $L^{-1}Z \sim N(0, I)$. From this we can then write down the log-likelihood, and numerically maximise.
- However, this procedure is (a) messy, (b) hard to generalise, (c) numerically unstable and (d) slow.
  - Note that it requires the inversion of a $T \times T$ matrix for each evaluation of the likelihood.
- The Kalman filter avoids both this mess, and the inversion of the $T \times T$ matrices.

The Kalman Filter: Assumptions

- Suppose x_t ("the state") is an n-dimensional stochastic process with $x_t = \mu_{t-1} + \mathrm{P}_{t-1}x_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim \mathrm{NIID}(0_n, \Sigma_{t-1})$.
  - Note that by augmenting the state space, any VARMA model may be placed in this form.
- Suppose that $x_0 \sim \mathrm{NID}(m_0, S_0)$, where m_0 and S_0 are known.
- Finally, suppose rather than observing x_t, we instead observe $y_t = \Gamma_{t-1}x_t + \eta_t$, where $\eta_t \sim \mathrm{NIID}(0, \Omega_{t-1})$.

- We assume throughout that μ_{t-1}, P_{t-1}, Σ_{t-1}, Γ_{t-1} and Ω_{t-1} are known in period $t-1$, and so independent of ε_t, η_t and x_t, but not necessarily independent of ε_{t-1}, η_{t-1} or x_{t-1}.
  - This allows for various classes of time-varying coefficient models, not least GARCH models.

A useful property of the Normal Distribution

- Suppose:

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix}, \begin{bmatrix} \Sigma_{UU} & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_{VV} \end{bmatrix} \right).$$

- Then:

$$U|V \sim \mathrm{N}(\mu_U + \Sigma_{UV}\Sigma_{VV}^{-1}(V - \mu_V), \Sigma_{UU} - \Sigma_{UV}\Sigma_{VV}^{-1}\Sigma_{VU}).$$

The Kalman Filter: One slide derivation!

- Let $\mathcal{F}_t$ denote all information available in period t.
  - $\mathcal{F}_t$ will contain $y_1, \ldots, y_t$, but not $x_1, \ldots, x_t$.
- Suppose $x_t|\mathcal{F}_t \sim \mathrm{N}(m_t, S_t)$.
- Then: $x_{t+1}|\mathcal{F}_t = \mu_t + \mathrm{P}_t x_t + \varepsilon_{t+1}|\mathcal{F}_t \sim \mathrm{N}(l_t, R_t)$, where $l_t \coloneqq \mu_t + \mathrm{P}_t m_t$ and $R_t \coloneqq \mathrm{P}_t S_t \mathrm{P}_t' + \Sigma_t$.
- So, since $y_{t+1} = \Gamma_t x_{t+1} + \eta_{t+1}$:

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} | \mathcal{F}_t \sim \mathrm{N}\left(\begin{bmatrix} l_t \\ \Gamma_t l_t \end{bmatrix}, \begin{bmatrix} R_t & R_t \Gamma_t' \\ \Gamma_t R_t & \Gamma_t R_t \Gamma_t' + \Omega_t \end{bmatrix} \right).$$

- Hence, by the previous result, if we define:

$$m_{t+1} \coloneqq l_t + R_t \Gamma_t' (\Gamma_t R_t \Gamma_t' + \Omega_t)^{-1} (y_{t+1} - \Gamma_t l_t), \text{ and}$$
$$S_{t+1} \coloneqq R_t - R_t \Gamma_t' (\Gamma_t R_t \Gamma_t' + \Omega_t)^{-1} \Gamma_t R_t,$$

- Then: $x_{t+1}|\mathcal{F}_{t+1} = x_{t+1}|y_{t+1}, \mathcal{F}_t \sim \mathrm{N}(m_{t+1}, S_{t+1})$.

The Kalman Filter: Initialization

- We often have no information about the initial values of our model's states, and so it is appropriate to initialize m_0 and S_0 to values reflecting this ignorance.
- This may be done by taking the limit as $k \to \infty$ of: $x_0|\mathcal{F}_{-k}$.
- In the absence of time varying coefficients, this will be distributed $\mathrm{N}(m_0, S_0)$, where $m_0 = \mu + \mathrm{P}m_0$ and $S_0 = \mathrm{P}S_0\mathrm{P}' + \Sigma$.

- When the model contains unit roots, such initialization is impossible (at least in the directions of the unit roots), and then it is more appropriate to use a highly diffuse initial distribution.

The Kalman Filter: "Kalman Smoothing"

- The Kalman Filter gives you an estimate of the state at time t conditional on observations in the current and previous periods.
  - These are often termed the "(Kalman) filtered states".
- However, future observations are also informative about the state at t.
  - For a simple example, imagine a model in which the state was observed with some number of periods of lag.
- If we are only interested in the states in the last few periods observed, then we may derive a full-information estimate of these states by working with an augmented set of state variables $\tilde{x}_t = [x_t' \quad \cdots \quad x_{t-l}']'$, much as we did when representing $\mathrm{VAR}(p)$ as $\mathrm{VAR}(1)$.
- More generally, to produce full-information estimates of the states in all periods, we have to combine the forward pass of the Kalman filter with a backwards pass which propagates the extra information from later observations back in time.
  - The most common is the Rauch–Tung–Striebel smoother.
  - Derivation is on the next slide.
  - The full information estimates of the states produced by such a smoother are termed the "(Kalman) smoothed states".

The Rauch–Tung–Striebel Smoother

- Recall $x_t|\mathcal{F}_t \sim \mathrm{N}(m_t, S_t)$, $x_{t+1} = \mu_t + \mathrm{P}_t x_t + \varepsilon_{t+1}$, $l_t \coloneqq \mu_t + \mathrm{P}_t m_t$ and $R_t \coloneqq \mathrm{P}_t S_t \mathrm{P}_t' + \Sigma_t$.
- Thus:

$$\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix} | \mathcal{F}_t \sim \mathrm{N}\left(\begin{bmatrix} m_t \\ l_t \end{bmatrix}, \begin{bmatrix} S_t & S_t \mathrm{P}_t' \\ \mathrm{P}_t S_t & R_t \end{bmatrix} \right).$$

- Hence by the Markovian property of x_t and the given property of Normal distributions:
  $x_t|x_{t+1}, \mathcal{F}_T = x_t|x_{t+1}, \mathcal{F}_t \sim \mathrm{N}(m_t + S_t \mathrm{P}_t' R_t^{-1}(x_{t+1} - l_t), S_t - S_t \mathrm{P}_t' R_t^{-1} \mathrm{P}_t S_t)$
- Now suppose: $x_{t+1}|\mathcal{F}_T \sim \mathrm{N}(a_{t+1}, B_{t+1})$.
- Then: $x_t|\mathcal{F}_T \sim \mathrm{N}(a_t, B_t)$, where:

$$a_t = m_t + S_t \mathrm{P}_t' R_t^{-1}(a_{t+1} - l_t), \text{ and}$$
$$B_t = S_t \mathrm{P}_t' R_t^{-1} B_{t+1} R_t^{-1} \mathrm{P}_t S_t + S_t - S_t \mathrm{P}_t' R_t^{-1} \mathrm{P}_t S_t.$$

Estimating state space models using the Kalman Filter

- The Kalman Filter gives us a "guess" of the unobserved state in the form of its distribution.
- The "prediction error decomposition" gives us a way of mapping this into a likelihood.
- Let θ collect all of the model's parameters (so μ_t, P_t, Σ_t, Γ_t and Ω_t are only functions of θ, m_t and S_t, for all t).
- Then, the prediction error decomposition just states that:

$$p(y_T, \dots, y_1|\theta) = p(y_T|y_{T-1}, \dots, y_1)p(y_{T-1}, \dots, y_1) = \cdots$$
$$= \prod_{t=1}^{T} p(y_t|\mathcal{F}_{t-1})\,.$$

- But from the derivation of the Kalman filter, we know $y_t|\mathcal{F}_{t-1} \sim \mathrm{N}(\Gamma_{t-1} l_{t-1}, \Gamma_{t-1} R_{t-1} \Gamma'_{t-1} + \Omega_{t-1})$, so this expression is easy to calculate.
  - We don't even need to calculate the smoothed states.
- We then just numerically maximise the likelihood.

Some practical considerations

- Although the Kalman Filter is orders of magnitude faster and more numerically robust than the crude method we suggested in our motivation, it still suffers from some numerical problems.
- In particular, due to the numerical quirks of matrix inversion and round-off error, the variant above is not guaranteed to always produce a positive definite S_t.
- As a result, the simple version here is not recommended.
- Instead, it is recommended to use one of the so called Square Root forms, which instead update some matrix decomposition of S_t (or its inverse).
  - There are many different varieties depending on which matrix decomposition they update.
  - The traditional Square Root Kalman Filter updates the Cholesky decomposition of S_t.
  - These forms all give the same results as the Kalman Filter when evaluated to infinite precision, but will perform much better at machine precision.

Macro applications of the Kalman Filter 1: Limited information models

- Some macro models incorporate limited information on the part of agents.
- For example, some models assume that the logarithm of labour productivity follows a process of the form: $a_t = z_t + d_t$, where $z_t = z_{t-1} + \varepsilon_t$ and $d_t = \rho d_{t-1} + \nu_t$. They further assume that only a_t is observable.
  - As a result, agents must run the Kalman Filter in order to predict their future productivity, and hence decide on investment levels.
  - This changes dynamics, since when they see an increase in a_t, they do not know if it is going to be permanent (in which case they should raise their capital stock by the same amount), or transitory (in which case they may hardly wish to adjust capital at all).
  - As a result, transitory shocks are given additional persistence, and investment will follow a hump-shaped path as agents realise the shock was not in fact permanent.
- Other models have assumed that the central bank's inflation target is determined by an unobservable stochastic process.
  - Again, agents will have to run the Kalman filter to determine how to react to monetary policy shocks.

Macro applications of the Kalman Filter 2: ML estimation of linearised models

- Once a DSGE model has been linearised, it and its observation equations are automatically in the form required for the Kalman Filter.
- As a result, we may use the Kalman Filter for ML estimation of linearised DSGE models.
  - This is what e.g. Dynare does.
- The methods we presented in the last topic for non-linear ML estimation would only handle the case in which there were as many shocks as observables.
- However, the Kalman filter has no problem with models in which there are many more shocks than observables, allowing for the ML estimation of models in which it is assumed all variables are observed with measurement error.
- In practice, however, the misspecification endemic in most DSGE models renders ML estimation a little unreliable.
- The likelihood is often multi-modal, and quite flat around the true maximum.
- This renders numerical maximisation difficult, and so too the computation of the likelihood's Hessian once a maxima is found.
- My recommendation is to use the CMA-ES algorithm (Hansen 2006), which has reasonable global search properties, but even then results should be taken cautiously.
  - Code etc. is here: https://www.lri.fr/~hansen/cmaesintro.html

Macro applications of the Kalman Filter 3: Bayesian estimation of linearised models

- Recall that Bayesian estimation is based on the identity:

$$\mathbb{E}[f(\theta)|X] = \int f(\theta)\frac{p(X|\theta)p(\theta)}{p(X)}d\theta\,.$$

- $p(X|\theta)$ is the likelihood.
- So just as the Kalman filter is the core of the standard ML estimation algorithm, its also the core of the standard Bayesian algorithm.

- Appropriate choice of priors may lessen some of the biases due to the misspecification of the underlying model.
- However, care should be taken to avoid using the same information twice.
  - The data set from which priors were produced should always be independent of that on which the data is estimated.
  - Too often in the literature this is not true.

Estimating non-linear models: Setup

- Suppose x_t is an n-dimensional stochastic process with $x_t = f_{t-1}(x_{t-1}, \varepsilon_t)$, where $\varepsilon_t \sim \mathrm{NIID}(0, \Sigma_{t-1})$.
  - The use of Gaussian shocks is WLOG, since we may convert the Gaussian distribution into any other via a suitable non-linear transform in f.
- Suppose that x_0 is a draw from some known distribution $p(x_0)$.
- Finally, suppose rather than observing x_t, we instead observe $y_t = g_{t-1}(x_t, \eta_t)$, where $\eta_t \sim \mathrm{NIID}(0, \Omega_{t-1})$.

- Much as before, the functions f_{t-1} and g_{t-1} are assumed to be known in period $t-1$, and hence independent of the period t shocks.

Separable non-linear form

- It is often much easier to work with models in which f and g are additively separable, so $f_{t-1}(x_{t-1}, \varepsilon_t) = f_{t-1}(x_{t-1}) + \varepsilon_t$ and $g_{t-1}(x_{t-1}, \eta_t) = g_{t-1}(x_{t-1}) + \eta_t$.
- Luckily, any non-separable model may be converted into a separable one by augmenting the state.

- In particular, define: $\tilde{x}_t = [x_t' \quad e_{1,t}' \quad e_{2,t}']'$, and replace $x_t = f_{t-1}(x_{t-1}, \varepsilon_t)$, $y_t = g_{t-1}(x_t, \eta_t)$, with:

$$x_t = f_{t-1}\big(x_{t-1}, e_{1,t-1}\big), \qquad y_t = g_{t-1}\big(x_t, e_{2,t-1}\big).$$

- Finally add new equations: $e_{1,t} = \nu_{1,t}$, $e_{2,t} = \nu_{2,t}$, where $\nu_{1,t}$ has the same distribution as ε_{t+1} and $\nu_{2,t}$ has the same distribution as η_{t+1}.

- Henceforth, we will assume separability.

The Extended Kalman Filter (EKF)

- The Extended Kalman Filter exploits two facts.
  1. The normal Kalman Filter is correct for linear models with time varying coefficients, even if those coefficients are a function of the lagged state.
  2. With small shocks, the behaviour of many models may be well approximated by a linear approximation around an estimate of the last value of the state.
- The EKF approximates the transition equation by its first order Taylor approximation around m_{t-1}.

$$x_t \approx f_{t-1}(m_{t-1}) + f'_{t-1}(m_{t-1})(x_{t-1} - m_{t-1}) + \varepsilon_t.$$

- Similarly, the EKF approximates the observation equation by:

$$y_t \approx g_{t-1}(m_{t-1}) + g'_{t-1}(m_{t-1})(x_{t-1} - m_{t-1}) + \eta_t.$$

- These approximated equations are linear, so the standard Kalman filter may then be applied to them. (m_0 is set to $\mathbb{E}x_0$ and S_0 is set to $\mathrm{Var}\, x_0$.)
- The approximation will be exact if $m_{t-1} = x_{t-1}$.
- A recent macro application of a similar local linearization technique to the estimation of a nonlinear model is Hall (2012).

The Kollman trick for non-linear DSGE models (1/2)

- Recently, Robert Kollman (2013) showed that the Kalman filter could be used for approximate filtering of pruned perturbation approximations.
- Recall that the second order perturbation approximation to a DSGE model takes the form:

$$x_t = A_0 + A_1 x_{t-1} + A_{11}(x_{t-1} \otimes x_{t-1}) + A_{12}(x_{t-1} \otimes \varepsilon_t) + A_2 \varepsilon_t + A_{22}(\varepsilon_t \otimes \varepsilon_t).$$

- However, the presence of the $x_{t-1} \otimes x_{t-1}$ term means this may be unstable, even if the BK conditions are satisfied for the linearised model.
- Thus Kim et al. (2008) suggested a "pruned" approximation was preferable. This takes the (augmented state space) form:

$$x_{1,t} = A_1 x_{1,t-1} + \varepsilon_t,$$
$$x_{2,t} = A_0 + A_1 x_{2,t-1} + A_{11}(x_{1,t-1} \otimes x_{t-1}) + A_{12}(x_{1,t-1} \otimes \varepsilon_t) + A_2 \varepsilon_t + A_{22}(\varepsilon_t \otimes \varepsilon_t).$$

- The initial justification was a little ad hoc, and met with some hostility, however this process has subsequently been put on much firmer theoretical ground by Lombardo (2010) and Lan and Meyer-Gohde (2013), who also extend the procedure to higher orders.
  - The Lan and Meyer-Gohde (2013) procedure seems to be the most accurate, but differences only become noticeable at high orders of approximation.
  - Dynare does not at present use this one.

The Kollman trick for non-linear DSGE models (2/2)

- The Kollman (2013) trick for estimating these pruned perturbation approximations is to approximate the joint distribution of ε_t and $\varepsilon_t \otimes \varepsilon_t$ by a Normal.
- The optimal approximation may be readily calculated analytically, given the current estimate of the covariance matrix of ε_t.
- Once this approximation is made, the system is linear in the augmented state vector: $[x'_{1,t} \quad x'_{1,t} \otimes x'_{1,t} \quad x_{2,t}]'$, with only the covariance matrix varying over time (thanks to the $x_{1,t-1} \otimes \varepsilon_t$ term).
- Thus, the standard Kalman filter may be applied.
- Obviously, by ignoring the relationship between $x_{1,t}$ and $x_{1,t} \otimes x_{1,t}$ in the state, we are losing some efficiency, however this cost is arguably worth paying to have a fast algorithm.

General framework for the estimation of nonlinear models: Prediction

- Suppose we know $p(x_t|\mathcal{F}_t)$. Then:

$$
\begin{aligned}
p(x_{t+1}|\mathcal{F}_t) &= \int p(x_{t+1}, x_t|\mathcal{F}_t)\, dx_t \\
&= \int p(x_{t+1}|x_t, \mathcal{F}_t) p(x_t|\mathcal{F}_t)\, dx_t \\
&= \int p(x_{t+1}|x_t) p(x_t|\mathcal{F}_t)\, dx_t \\
&= \int p_{\mathrm{N}}(x_{t+1}|f_t(x_t), \Sigma_t) p(x_t|\mathcal{F}_t)\, dx_t\,,
\end{aligned}
$$

  - where $p_{\mathrm{N}}(U|\mu_U, \Sigma_U)$ is the p.d.f. of a Normal distribution with mean μ_U and variance Σ_U.

General framework for the estimation of nonlinear models: Update

- By Bayes' Rule:

$$
\begin{aligned}
p(x_{t+1}|\mathcal{F}_{t+1}) = p(x_{t+1}|y_{t+1},\mathcal{F}_t) &= \frac{p(x_{t+1},y_{t+1}|\mathcal{F}_t)}{p(y_{t+1}|\mathcal{F}_t)} \\
&= \frac{p(y_{t+1}|x_{t+1},\mathcal{F}_t)p(x_{t+1}|\mathcal{F}_t)}{p(y_{t+1}|\mathcal{F}_t)} \\
&= \frac{p(y_{t+1}|x_{t+1}\,)p(x_{t+1}|\mathcal{F}_t)}{p(y_{t+1}|\mathcal{F}_t)} \\
&= \frac{p_{\mathrm{N}}(y_{t+1}|g_t(x_t),\Omega_t)p(x_{t+1}|\mathcal{F}_t)}{p(y_{t+1}|\mathcal{F}_t)} \\
&\propto p_{\mathrm{N}}(y_{t+1}|g_t(x_t),\Omega_t)p(x_{t+1}|\mathcal{F}_t).
\end{aligned}
$$

- $p(x_{t+1}|\mathcal{F}_t)$ was calculated on the previous slide, thus, given $p(x_t|\mathcal{F}_t)$ we can work out $p(x_{t+1}|\mathcal{F}_{t+1})$.
- However, this requires the evaluation of a rather nasty integral! It also requires holding full distributions in memory, not just their mean and covariance. These are difficult problems, which we will discuss next week.