14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 27, 2007
revised November 30, 2009

Lecture 22

State-space models. ML estimation. DSGE models.

Examples of State-space models (cont.)

Missing or unequally spaced observations

Suppose we have a state-space model where we are missing observations. The model is:

$$\begin{aligned} y_t =& z_t\alpha_t + S_t\xi_t \\ \alpha_t =& T_t\alpha_{t-1} + R_t\eta_t \end{aligned}$$

But instead of observing all $\{y_t\}_{t=1}^T$, we only observe some subset $\{y_{i1},..,y_{i\tau}\} = \{y_i | i \in I\}$. To produce a likelihood we can consider the following model:

$$\begin{aligned} y_t^* =& z_t^*\alpha_t + w_t^* \\ \alpha_t =& T_t\alpha_{t-1} + R_t\eta_t \end{aligned}$$

where $y_t^* = \begin{cases} y_t & t \in I \\ N_t & \text{otherwise} \end{cases}$, $z_t^* = \begin{cases} z_t & t \in I \\ 0 & \text{otherwise} \end{cases}$, and $w_t^* = \begin{cases} w_t & t \in I \\ N_t & \text{otherwise} \end{cases}$, with $N_t \sim iid\, N(0,1)$. Assume that we observe $y_t^* = 0$ when $t \notin I$, then the conditional likelihood is

$$f(y_t^*|\mathcal{Y}_{t-1}^*) = \begin{cases} \phi(0) & t \notin I \\ f(y_t|\{y_s : s \in I,\, s < t\}) \end{cases}$$

and the likelihood is $f(y_1^*,...,y_T^*) = \phi(0)^{\#I^c} f(y_{i1},...,y_{i\tau}) = (2\pi)^{-\frac{\#I_c}{2}} f(y_{i1},...,y_{i\tau})$. It means that we can run Kalman filter for y_t^*, it produces(up to a constant) a valid likelihood for the initial model.

Example 1. Suppose:

$$y_t = \phi y_{t-1} + \epsilon_t$$

and we observe $t \in I = \{1,3,4,5\}$. In state space form, as above, we have:

$$\begin{aligned} \alpha_t &= \alpha_{t-1}\phi + \epsilon_t \\ y_t^* &= z_t\alpha_t + w_t \end{aligned}$$

where $z_t = \begin{cases} 1 & t \in I \\ 0 & \text{otherwise} \end{cases}$ and $w_t = \begin{cases} 0 & t \in I \\ N(0,1) & \text{otherwise} \end{cases}$. Let's think what Kalman filter would do: for $t \in I$ we observe $y_t = \alpha_t$, so our best linear predictor of α_t is y_t. For $t = 2$, y_2 is unrelated to α_2, so our best linear predictor of α_2 is just ϕy_1. Then the conditional means used in the Kalman filter are:

$$\alpha_{t|t} = \begin{cases} y_t & t \in I \\ \phi y_{t-1} & t = 2 \end{cases}$$

To form the conditional likelihood, we need the distribution of $y_t|\mathcal{Y}_{t-1}$, which has mean

$$y^*_{t|t-1} = \phi\alpha_{t-1|t-1} = \begin{cases} \phi y_{t-1} & t \in \{4,5\} \\ \phi^2 y_1 & t = 3 \end{cases}$$

and variance

$$F_t = \begin{cases} \sigma^2 & t \in \{4,5\} \\ (1+\phi^2)\sigma^2 & t = 3 \end{cases}$$

We only need the conditional distribution at $t = 3, 4, 5$ because the likelihood is:

$$f(y_1, y_3, y_4, y_5) = f(y_1)f(y_3|y_1)f(y_4|y_3, y_1)f(y_5|y_4, y_3, y_1)$$

and the conditional (on y_1) likelihood is

$$\begin{aligned} f(y_1, y_3, y_4, y_5|y_1) =& f(y_3|y_1)f(y_4|y_3, y_1)f(y_5|y_4, y_3, y_1) \\ =& \frac{1}{\sigma\sqrt{1+\phi^2}}\phi\left(\frac{y_3 - \phi^2 y_1}{\sigma\sqrt{1+\phi^2}}\right)\frac{1}{\sigma}\phi\left(\frac{y_4 - \phi y_3}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{y_5 - \phi y_4}{\sigma}\right) \end{aligned}$$

where $\phi(\cdot)$ is the normal pdf.

The nice thing is that Kalman filter does this reasoning automatically.

Loose end: Initial Value

The Kalman filter depends on unknown initial values, $\alpha_{1|0}$ and $P_{1|0}$. Recall that these are the mean and variance of α_1 given observations up to time 0. The model is

$$\begin{aligned} y_t =& z_t\alpha_t + S_t\xi_t \\ \alpha_t =& T_t\alpha_{t-1} + R_t\eta_t \end{aligned}$$

One way to approach the problem is to assume (whenever it is possible) that α_t is a stationary process. It is possible if $T_t = T$ and $|\text{eigenvalues of } T| < 1$, then the we consider α_1 to be distributed according to invariant distribution: normally with stationary mean 0 and unconditional stationary variance P satisfying:

$$P = TPT' + RQR'$$

so

$$vec(P) = [I - T \otimes T]^{-1} vec(RQR')$$

.

Non-stationary Two approaches:

1. Diffuse prior: $\alpha_{1|0} = 0$ and $P_{1|0} = kI$, where k is large.

   Example 2.

   $$\begin{aligned} y_t =& \mu_t + \epsilon_t \\ \mu_t =& \mu_{t-1} + \eta_t \end{aligned}$$

   where $\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \sim iidN(0, \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix})$. Then the updating is:

   $$\mu_{t+1|t} = \mu_{t|t} = (1 - k_t)\mu_{t|t-1} + k_t y_t$$

where $k_t = \frac{P_{t|t-1}}{P_{t|t-1}+1}$ and

$$P_{t+1|t} = P_{t|t-1} - \frac{P_{t|t-1}^2}{1+P_{t|t-1}} + q$$

Then, for example,

$$\mu_{2|1} = [1 - \frac{k}{1+k}]\mu_{1|0} + \frac{k}{1+k}y_1 \to y_1 \text{ as } k \to \infty$$

Similarly, $P_{2|1} \to 1+q$. That is, the uninformative(diffuse) prior is equivalent to a prior concentrated on the first observation.

2. Treat $\alpha_{1|0}$ as a parameter to estimate and $P_{1|0} = 0$

Maximum Likelihood Estimation (MLE)

One use of Kalman filter is to calculate Likelihood and find ML estimator. However, the standard MLE asymptotic theory is stated in iid case. Under what conditions MLE is consistent and asymptotically normal? As always, there is a great variety of cases, we will discuss one- with stationary state variable. The conditions roughly will be:

1. Model should be identified. These is obviously a necessary condition, but it is not always trivial to verify. For example, we know that MA models are locally but not globally identified, and we do not know how to check for global identification
2. Stationarity and |eigenvalues of $T| < 1$
3. θ_0 is not on a boundary

Under these three conditions,

$$\sqrt{T}(\hat{\theta} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0))$$

where $I(\theta_0) = \lim \frac{1}{T} E \sum_t \frac{\partial^2 f}{\partial\theta\partial\theta'}(y_t|\mathcal{Y}_{t-1};\theta)$ is the information matrix. Note that this is not just a restatement of the usual MLE consistency and asymptotic normality result, because in this case we do not have iid data. The key for this result is condition 2. It implies that the Kalman filter converges as time goes to infinity, in the sense that $F_t \to F$, a constant matrix, and $P_{t|t-1} \to P$. To show this, you would write down the Riccati equation, which gives $P_{t+1|t}$ as a function of $P_{t|t-1}$. This equation converges exponentially fast.

Even without stationarity, MLE can work in some cases. However, think about unit root case: OLS is MLE under normality assumptions. We know that it behaves in non-classical way!

Quasi-ML

Kalman filter makes some strong assumptions. In particular, it assumes linearity and normal errors. Linearity is usually considered a reasonable assumption, but normality is not. This begs the question: what if we use ML to estimate θ, but the errors are not really normal? This situation is called quasi-ML. It was originally analyzed by White (1982) for iid case.

Let's concentrate on iid case ourselves. Assume that one has a sample $x_t \sim i.i.d.\ g(\cdot)$, where g is the pdf. However, s/he runs misspecified MLE faulty assuming $x_t \sim f(x;\theta)$.

Under some technical assumptions, the estimate is consistent for the parameter θ^* (called pseudo-parameter) that minimizes the KLIC (Kullback-Leibler Information Criterion).

$$KLIC(g, f; \theta) = E\left(\log\left[\frac{g(x_t)}{f(x_t;\theta)}\right]\right)$$

The MLE is asymptotically normal, but the variance-covariance matrix should be corrected. To see this consider the FOC from maximizing the likelihood:

$$0 = \sum_t \frac{\partial}{\partial\theta} \log f(x_t; \hat{\theta})$$

Take a Taylor expansion around θ^*:

$$0 = \sum_t \frac{\partial}{\partial\theta} \log f(x_t; \theta^*) + (\hat{\theta} - \theta^*) \sum \frac{\partial^2}{\partial\theta^2} \log f(x_t; \theta^*) + o_p(1)$$

$$\sqrt{T}(\hat{\theta} - \theta_0) = \frac{\frac{1}{\sqrt{T}} \sum_t \frac{\partial}{\partial\theta} \log f(x_t; \theta^*)}{\frac{1}{T} \sum \frac{\partial^2}{\partial\theta^2} \log f(x_t; \theta^*)} \Rightarrow \frac{N(0, I_1)}{I_2}$$

where $I_1 = \lim \frac{1}{T} \sum \left[\frac{\partial}{\partial\theta} \log f(x_t, \theta^*)\right]' \left[\frac{\partial}{\partial\theta} \log f(x_t, \theta^*)\right]$ and $I_2 = \lim \frac{1}{T} E\left[\sum \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_t, \theta^*)\right]$. If the likelihood were correctly specified, then the information equality would hold and $I_1 = I_2$. However, this is not true when the likelihood is misspecified, so

$$\sqrt{T}(\hat{\theta} - \theta_0) \Rightarrow N(0, (I_2 I_1^{-1} I_2)^{-1})$$

It is considered a good econometric practice to use this “sandwich” form for estimation of standard errors.

DSGE vs VAR

Nowadays there are two competing approaches to empirical macroeconomics: VAR and DSGE. The initial idea behind the VARs is to try to make as few assumptions as possible. Identification is achieved by imposing assumptions on which several competing theories would agree. We have discussed that, in fact, there are many implicit assumptions behind VARs such as linearity and fundamentalness of the shocks. DSGE places more structure on the model starting from micro-foundations.

There are serious (real-world) empirical questions that can be addressed by both VARs and DSGE, such as:

- Ask quantitative counterfactual policy questions
- Make conditional forecasts
- Examine effects of past policy changes (effects on means and variances, e.g. Great Moderation debate)

But there are questions that can be handled only by structural models such as DSGE:

- Examine a welfare effect of different policies
- Solve for optimal policies

The relation between VARs and DSGEs are not fully clear, an important paper on the topic is Fernandez-Villaverde, Rubio-Ramirez, Sargent “ A,B,CS (and D)S for understanding VARs”

I would divide the way to approach DSGE into “Macro” and “Econometrics” part. The division is not exact, I put everything I don’t want to talk about into “Macro” part:) The main focus is on “linearized models”; non-linear solutions/models are more difficult and less developed (if interested, go to Jesus Fernandez-Villaverde web-site). Linear models can utilize Kalman filters, while non-linear would go for partical filtering, which we have not discussed.

"Macro" Steps :

1. Write down (non-linear optimization) model :
    - Who maximizes what?
    - balances
2. Solve model
    - Euler equations and other first order conditions
    - Usually log-linearize around non-stochastic steady-state
    - then solve rational expectations to get model in state-space form $\begin{cases} \alpha_t = & T(\theta)\alpha_{t-1} + \epsilon_t \\ y_t = & z(\theta)\alpha_t + v_t \end{cases}$ where coefficients ($T(\theta)$ and $Z(\theta)$) are functions of the structural parameters (θ), usually $T()$ and $Z()$ are highly nonlinear and calculated numerically

"Macro" Steps : There are 3 methods for estimation: Indirect Inference (other terms: Simmulated GMM, calibration, matching impulse responses(auto-covariances, moments, etc)); Maximum Likelihood; Bayesian. The first we discussed several lectures ago, the last will be discussed in future. Today our focus on ML.

1. Given the state-space representation, write down likelihood function (use Kalman filter)
2. ML: $\hat{\theta} = \arg\max l(\theta)$, compute $s.e.(\hat{\theta})$ (use sandwhich formula in order to robustify to misspecification of distribution functions)

Papers by Ireland explicate this procedure.

Potential difficulties of MLE for DSGE

1. Stochastic Singularities

DSGE carries nontrivial implications only for the number of series equal to the number of shocks.

Example 3. *Ingram, Kocherlakota, Savin (1994)*: Suppose we observe consumption and production, $\{c_t, y_t\}$ generated from the following model:

$$\max_{c_t} E_0 \sum_t \beta^t \ln c_t$$
$$\text{s.t. } c_t + I_t = y_t = A_t k_t$$
$$k_{t+1} = (1-\delta)k_t + I_t$$

where A_t is a productivity shock (the only shock in this economy). We observe two series, $\{c_t, y_t\}$, but only have one shock. That would imply some deterministic relation between two observed series. Let's solve the model. Due to log-utility the consumer has a myopic behavior- he consumes a fixed fraction $(1-\beta)$ of all available resources, and put everything else to future capital:

$$\begin{aligned} c_t =&(1-\beta)(1-\delta+A_t)k_t \\ k_{t+1} =&\beta(1-\delta+A_t)k_t \end{aligned}$$

or in terms of y and c

$$\begin{aligned} y_{t+1} =&A_{t+1}\frac{\beta}{1-\beta}c_t \\ c_t =&\beta c_{t-1}(1-\delta+A_t) \end{aligned}$$

so

$$c_t = \frac{\beta^2}{1-\beta} y_t + \beta(1-\delta)c_{t-1}$$

there is a deterministic relationship between c_t and y_t. This relationship will almost never hold in the data. If you run ML, you will face a problem, since data rejects the model.

More generally, whenever a model has more observable series than shocks, there will be a deterministic relationship among the observed variables, which will not hold in the data. There are two solutions:

1. Define as many shocks as variables (Ingram, Kocherlakota, Savin (1994)). In example above, one may assume that δ_t is a random rate of depreciation to add another shock to the system. This is a "creative way"- you have to label many "shocks".

2. Assume measurement errors. $Y_t = Z(\theta)\alpha_t + U_t$, here Y_t is observable, U_t is a measurement error (usually the absence of cross- and auto- correlations is assumed).

   Note There is a fundamental problem (Prescott'86): All models are necessarily highly abstract (necessary statistically rejected by the data). Adding measurement error does not necessary reduce misspecification: ML (as opposed to GMM) is a *full information* method, i.e., produces *joint estimation* of all relationships imposed by the model. In GMM you can choose which relations to impose. Additional question is a choice of observables (smaller number require smaller set of errors). A new and quite nice idea is to put factor structure in the model: for example, one does not observe inflation *per se* but rather several noisy measures of inflation- GDP deflator, consumer price index, etc. For an example, see Boivin and Giannoni(2005).

2. Identification

As you add more shocks, you add more parameters to the model, so you might lose identification. Also, if we solve the model by linearizing and estimating the state-space form, we do not know if the MA model is globally identified. Moreover, we do not know the nature of the mapping from the state-space MA coefficients to the structural parameters. This mapping is difficult to analyze because it is computed numerically.

Example 4. *Canova & Sala*:

$$\begin{aligned} \max \sum_t \beta^t \frac{c_t^{1-\phi}}{1-\phi} \\ \text{s.t.} c_t + k_{t+1} = A_t k_t^{1-\eta} + (1-\delta)k_t \\ A_t = \rho A_{t-1} + \mu + \epsilon_t \end{aligned}$$

Canova and Sala (see handouts) simulated this model and found that likelihood is nearly flat along some manifolds in the parameter space. Is it a problem? No clear answer.

How we define the lack of identification? Observational equivalence. It seems that in the example above there is no observational equivalence (the expected likelihood at different values is different), thus, we are identified in a classical sense. I am aware of a single paper that directly addresses the identification issue- Comunjer and Ng(2009)

What about weak identification? The term is not well-defined in the literature so far. We would define a "weak identification" as a situation in which inferences based on the MLE asymptotic theory are misleading, that is, the ML estimates are very biased, the t-tests based on normal approximation have bad size properties, confidence sets have low coverage, etc.

As a demonstration of such a situation we(Jim Stock and me) took "the simplest possible" DSGE model:

$$\begin{array}{cc} \beta E_t \pi_{t+1} + \kappa x_t - \pi_t + \varepsilon_t = 0 & \text{Calvo pricing} \\ -[r_t - E_t \pi_{t+1} - rr_t^*] + E_t x_{t+1} - x_t = 0 & \text{intertemporal equation} \\ r_t = \alpha r_{t-1} + (1-\alpha)\phi_\pi \pi_t + (1-\alpha)\phi_x x_t + u_t & \text{policy rule} \\ rr_t^* = \rho \Delta a_t & \text{definition of natural rule} \end{array}$$

The model has 3 observable variables : x_t- output gap, r_t- real interest rate, π_t- inflation; one latent variable: rr_t^*- natural rate; and 3 exogenous shocks: $\Delta a_t = \rho \Delta a_{t-1} + \varepsilon_a$- technology shock, $u_t = \delta u_{t-1} + \varepsilon_u$- policy shock, and ε-mark-up shock.

We study the model in the neighborhood of the following parameter values:Parameters: discount factor $\beta = 0.99$; policy parameters $\phi_x = 0, \phi_\pi = 1.5, \alpha = 0$; Calvo parameter $\kappa = \frac{(1-\theta)(1+\phi)(1-\beta\theta)}{\theta}, \theta = 0.75$; persistence of technology shock $\rho = 0.2$; persistence of policy shock $\delta = 0.2$; $\sigma_a^2 = Var(\varepsilon_a) = 1, \sigma_u = Var(\varepsilon_u) = 1, \sigma = Var(\varepsilon) = 1$. We treat β as known, since it is difficult to estimate, and in the majority of papers it is calibrated. We treat all other 9 parameters as unknown.

We simulated a data sets from the model and then used them as "data" in estimating the model using MLE. Part of the results are below, the rest are in handouts.

	ϕ_x	ϕ_π	α	ρ	δ	κ	σ_a	σ_u	σ
bias	0.4407	0.7124	0.0015	-0.0031	0	0.1305	0.3637	0.5753	0.1803
sqrt(MSE)	1.0645	1.6147	0.0771	0.0853	0.0329	0.3904	0.998	1.3424	0.5361
size of t-test	0.46	0.26	0.2	0.14	0.14	0.6	0.34	0.72	0.64
abnormal t	0.14	0.1	0.02	0.04	0	0.44	0.18	0.36	0.26
p-value KS(t)	0	0	0.01	>0.5	>0.5	0	0	0	0
size of LR	0	0	0.08	0.08	0	0.04	0.08	0	0.02
p-value KS (LR)	0.0497	0.0266	0.7036	0.0608	0.7381	0.2584	0.1434	0.0999	0.2576

Explanation: the first block of the results describe the biases of MLE (the comparable scale is MSE). The second block is about the behavior of t-statistics(for the corresponding coefficients). One can see that the real size of t-tests(5% nominal level) can be as high as 60% in some cases. Line "abnormal t" stays for the fraction of t-statistics which exceed 10 in absolute value. This line shows that the finite sample distribution of t-stat has thicker tails than prescribed by the standard normal. The line "p-value KS(t)" reports the result of Kolmogorov-Smirnoff test that the finite sample t-statistic is normally distributed. The third block of results describe the behavior of LR test for the value of one particular coefficient. LR statistic is calculated as the doubled difference between full log-likelihood and a concentrated log-likelihood(when the corresponding coefficient is concentrated out). According to ML asymptotic theory, it is supposed to be χ_1^2-distributed. We can see that LR test is too conservative(lacking power), and does not have χ^2- distribution in finite samples.

The described results seem to be very typical for weak instruments literature.

One may ask, whether you can detect this "weak identification" in your application by looking at likelihood function. Say, should you expect flat or bumpy likelihood? Not obligatory. It is difficult to visualize the multi-dimensional function. We draw one-dimensional and two-dimensional cuts of the typical likelihood, and it looks regular (it looks quadratic).

3. Optimization

Optimization in many dimensions is difficult because it is hard to visualize and many things can go wrong. The objective function might have local extrema, discontinuities, and be non-differentiable. Because of the objective function might be discontinuous and non-differentiable, many people think you should use a derivative free optimization algorithm. A popular method is the simplex method (sometimes called the Nelder-Mead simplex algorithm to avoid confusion with the simplex algorithm from linear programming). It is easiest to explain and visualize in two dimensions. Wikipedia has a nice animation illustrating the algorithm (http://en.wikipedia.org/wiki/Nelder-Mead_method). We have a function $f : \Re^n \to \Re$ that we want to minimize. We start with $n+1$ (which form a simplex in $\Re^n$). We evaluate the function at each of the points. We then try to improve the worse function value by replacing the worst point. We do some combination of reflecting, expanding, and contracting the simplex to do this.

The simplex will converge to a local extremum. In practice, it is necessary to begin the algorithm from multiple starting values to gain confidence that you have found the global extremum.

General suggestions :

- Start optimization from different points

- Try multiple algorithms, both derivative and non-derivative based methods
- Try to visualize optimization results in many different dimensions
- Simulate model with estimated coefficients and study identification