NBER Summer Institute Minicourse –
What's New in Econometrics: Time Series

Lecture 6:

July 15, 2008

Specification and estimation of models with stochastic time variation

Outline

1. Break Models (Break Dates)

2. Markov Switching Models

3. Martingale TVP

    a. MLEs and alternatives

    b. Data Augmentation (EM)

    c. TVPs as nuisance parameters

1. Break Models

Inference about Break Dates (Bai (1997), Hansen (2001))

Example (a special case of the linear regression):

$$y_t = \beta_t + \varepsilon_t \qquad \beta_t = \begin{cases} \beta \text{ for } t \leq \tau \\ \beta + \delta \text{ for } t > \tau \end{cases}$$

$\pi = \tau/T$ = Break "Fraction"

τ_{o} = true break date

π_{o} = true break fraction

$\hat{\tau}$ = Least squares estimator of τ, $\hat{\pi} = \hat{\tau}/T$

Some results that are useful for inference:

Bai(1997) shows $\hat{\pi} - \pi_{\text{o}} \sim O_p(T^{-1}\delta^{-2})$, so that

$$T\delta^2(\hat{\pi} - \pi_{\text{o}}) \sim O_p(1)$$

$$\delta^2(\hat{\tau} - \tau_o) \sim O_p(1)$$

Thus, $\hat{\pi}$ is consistent for π_{o}, but $\hat{\tau}$ is not consistent for τ_o.

The speed at which $\hat{\pi}$ converges to π_{o} depends on δ.

In general, the distribution of $\hat{\pi}$ and $\hat{\tau}$ depends on the distribution of the errrors ε_t. This is true even when T is large. Thus, robust inference is problematic.

There are approximations that can be used when δ is appropriately small.

An asymptotic approximation: Recall $T\delta^2(\hat{\pi}-\pi_o) \sim O_p(1)$, so assume $T\delta^2 \to\infty$. Also, δ is small, so assume $\delta \to 0$. (More formally, $\delta = \delta_T$ which approaches zero as T grows large.). (Example, both of these are satisfied if $\delta_T = aT^{-0.49}$)

The challenge is to compute a convenient expression $\hat{\pi}$. The trick is to use empirical process methods like those used for the FCLT. The main ideas can be understood in a situation in which β and δ are known, so that estimating τ (equivalently π) is the only problem.

The least squares objective function is

$$\text{SSR}(\tau) = \sum_{t=1}^{\tau}(y_t-\beta)^2 + \sum_{t=\tau+1}^{T}(y_t-\beta-\delta)^2$$

and the trick is to study the behavior of this function as T gets large.

Suppose $\tau > \tau_o$. Then we can write

$$SSR(\tau) = \sum_{t=1}^{\tau}(y_t - \beta)^2 + \sum_{t=\tau+1}^{T}(y_t - \beta - \delta)^2$$

$$= \sum_{t=1}^{\tau_o}(y_t - \beta)^2 + \sum_{t=\tau+1}^{T}(y_t - \beta - \delta)^2 + \sum_{t=\tau_o+1}^{\tau}((y_t - \beta - \delta) + \delta)^2$$

$$= \sum_{t=1}^{\tau_o}\varepsilon_t^2 + \sum_{t=\tau+1}^{T}\varepsilon_t^2 + \sum_{t=\tau_o+1}^{\tau}(\varepsilon_t + \delta)^2$$

$$= \sum_{t=1}^{T}\varepsilon_t^2 + (\tau - \tau_o)\delta^2 + 2\delta\sum_{t=\tau_o+1}^{\tau}\varepsilon_t$$

$$= \sum_{t=1}^{T}\varepsilon_t^2 + (\pi - \pi_o)T\delta^2 + 2\delta\sum_{t=1}^{[T(\pi-\pi_o)]}\varepsilon_{[\pi_o T]+1}$$

Where the last expression substitures $\pi = \tau/T$. The first term does depend on τ; ignore it when thinking about the function that is being maximizing.

Thus, we can think about choosing τ or π to minimize

$$(\tau - \tau_o)\delta^2 + 2\delta \sum_{t=\tau_o+1}^{\tau} \varepsilon_t = (\pi - \pi_o)T\delta^2 + 2\delta \sum_{t=1}^{[T(\pi-\pi_o)]} \varepsilon_{[\pi_o T]+1}$$

Let $\upsilon = (\pi - \pi_o)T\delta^2/\sigma_\varepsilon^2$.

Then minimizing *SSR* over τ is the same as minimizing g_T over υ, where

$$g_T(\upsilon) = \upsilon + 2(\delta/\sigma_\varepsilon) \sum_{t=1}^{[(\delta/\sigma_\varepsilon)^{-2}\upsilon]} (\varepsilon_{[\pi_o T]+1} / \sigma_\varepsilon)$$

and the division by σ_ε is for later convenience.

Recall $\delta \rightarrow 0$, so $\delta^{-2} \rightarrow \infty$, so that $(\delta / \sigma_{\varepsilon}) \sum_{t=1}^{[(\delta/\sigma_{\varepsilon})^{-2} \upsilon]} (\varepsilon_{[\pi_o T]+1} / \sigma_{\varepsilon}) \xrightarrow{d} W(\upsilon)$.

(For analogy with the standard formula, think of $(\delta/\sigma_{\varepsilon})^{-2} =$ *sample size*, so that $(\delta/\sigma_{\varepsilon}) = \frac{1}{\sqrt{sample\ size}}$).

Thus $g_T(\upsilon) = \upsilon + 2(\delta/\sigma_{\varepsilon}) \sum_{t=1}^{[(\delta/\sigma_{\varepsilon})^{-2} \upsilon]} (\varepsilon_{[\pi_o T]+1} / \sigma_{\varepsilon}) \xrightarrow{d} g(\upsilon) = \upsilon + 2W(\upsilon)$.

and from arguments like those used for the FCLT, $g_T(\,.\,) \Rightarrow g(\,.\,)$.

The argument for $\tau < \tau_o$ is similar. Putting these together, the least squares problem for estimating τ (or $\pi = \tau/T$) is approximately the same as

$$\min_{\upsilon} G(\upsilon) \text{ where } G(\upsilon) = \begin{cases} |\upsilon| + 2W_1(\upsilon) \text{ for } \upsilon \geq 0 \\ |\upsilon| + 2W_2(-\upsilon) \text{ for } \upsilon < 0 \end{cases}$$

where W_1 and W_2 are independent Wiener processes.

The value of υ that minimizes this random function has a very non-gaussian shape.

Here are some values

	c	
Probability	$\Pr(\lvert \hat{\upsilon} \rvert < c)$	Standard Normal Prob(|z| < c)
50%	2.8	
67%	4.4	1.0
80%	6.7	
90%	10.0	1.64
95%	13.8	1.96
99%	23.5	2.56

Constructing a confidence interval for τ: Ingredients, (T, $\hat{\sigma}_\varepsilon$, $\hat{\delta}$, $\hat{\pi}$ (or $\hat{\tau}$)):

We know (from the table on the last page) Pr($|\hat{\upsilon}|$ < 4.4) = 0.67. Using $\hat{\upsilon} = \frac{1}{\hat{\sigma}_\varepsilon^2} T\hat{\delta}^2(\hat{\pi} - \pi_o)$, a 67% confidence interval for π satisfies

$$\left|\frac{1}{\hat{\sigma}_\varepsilon^2} T\hat{\delta}^2(\hat{\pi} - \pi)\right| < 4.4 \text{ or } \hat{\pi} - 4.4\frac{\hat{\sigma}_\varepsilon^2}{T\hat{\delta}^2} < \pi < \hat{\pi} + 4.4\frac{\hat{\sigma}_\varepsilon^2}{T\hat{\delta}^2}$$

so that a 67% CI for τ is: $\hat{\tau} - 4.4\frac{\hat{\sigma}_\varepsilon^2}{\hat{\delta}^2} < \tau < \hat{\tau} + 4.4\frac{\hat{\sigma}_\varepsilon^2}{\hat{\delta}^2}$

(Programs: Bruce Hansen's webpage)

Empirical Example … Great Moderation

$\phi_t(\mathrm{L})\Delta y_t = \mu_t + \varepsilon_t \quad (Y = \ln(\mathrm{GDP}))$

$\varepsilon_t^2 = \sigma_t^2 + (\varepsilon_t^2 - \sigma_t^2)$

(Numbers from SW(2002))
p-value for break in $\phi_t(\mathrm{L})$ and $\mu_t = 0.98$
p-value for break in $\sigma_t^2 = 0.00$ ($\hat{\tau} = 1983:2$)

67% CI for Break in σ_t^2: 1982:4 – 1985:3

(What do you think about this confidence interval, given derivation of "Bai" confidence intervals for break dates?)

Multiple Deterministic Breaks: Bai and Perron (1998)

Single Joint Deterministic Breaks in Multiple Processes: Bai, Lumsdaine, Stock (1998)

Multiple Joint Deterministic Breaks in Multiple Processes; Qu and Perron (2007)

Using "Breaks": Historical Analysis vs. Forecasting

Stochastic Breaks: Markov Switching

(a) A 2-state version of Hamilton's Markov-Switching Model:

$$y_t = \mu(s_t) + \sigma(s_t)\varepsilon_t,\ \ s_t = 0 \text{ or } 1 \text{ with } P(s_t = i \mid s_{t-1} = j) = p_{ij}$$

Rewrite as $y_t = \mu_0(1-s_t) + \mu_1 s_t + \{\sigma_0(1-s_t) + \sigma_1 s_t\}\,\varepsilon_t$

Issues:

(i) Filtering and Smoothing given parameters

(ii) Testing for Markov Switching:

$$H_o: \mu_1 = \mu_0 \text{ and } \sigma_1 = \sigma_0$$

p_{ij}'s are unidentified (Andrews-Ploberger (1994), Hansen (1992))

(iii) Estimation: MLE (easy via data augmentation/EM, discussed below)

(v) Lots of extensions (use your imagination)

(iv) Changes are "recurrent"

(b) A non-recurrent model (inspired by Pesaran, Pettenuzzo, and Timmerman (2006))

Motivation: (i) Non-recurrent model; (ii) Deterministic break model is not useful for forecasting (it says nothing about post-sample breaks).

$$y_t = \mu(s_t) + \sigma(s_t)\varepsilon_t$$

K states: $s_t = 1,2,\ldots, K$

State Dynamics:

Initial condition: $s_1 = 1$.

At other dates there are two possibilities: $s_t = \begin{cases} s_{t-1} \text{ with prob } p \\ s_{t-1}+1 \text{ with prob } 1-p \end{cases}$

$\mu(s_t) = \mu + \eta(s_t)$ where $\eta(s_t) \sim \text{N}(0, \sigma_\eta^2)$ (similar for $\sigma(s_t)$)

Or, perhaps $\mu(s_t) = \mu(s_t - 1) + \eta(s_t)$
(PTT use hierarchical Bayes method for estimation)

Other ways of incorporating instability into forecasting models:

Add factors and intercept shifts:

(i) Subjective/Judgment

(ii) Differencing (e.g., Clements and Hendry (1999))

(iii) Martingale variation in intercept: Cooley and Prescott (1973a, 1973b, 1976)

Martingale Variation: Linear Models, say $y_t = \beta_t' x_t + \varepsilon_t$ with $\beta_t = \beta_{t-1} + \eta_t$. (Textbook References: Hamilton (1994), Harvey (1989).)

Running example: $y_t = \beta_t + \varepsilon_t$

$$\beta_t = \beta_{t-1} + \eta_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix} \right)$$

Estimation Issues: Parameters (β_0, σ_ε, and σ_η)

β_0: Initialize Kalman Filter with "Vague" prior $\beta_{0/0} \sim N(0, \kappa)$, where $\kappa \approx \infty$. Then $\beta_{0/T}$ is the GLS (Gaussian MLE) estimator of β_0 (give σ_ε and σ_η).

σ_ε and σ_η: Nonlinear maximization of log-likelihood (which can be computed using Kalman Filter as described in Lecture 5).

2 issues: Problems with MLE when $\sigma_\eta / \sigma_\varepsilon$ is small. Computational problems in large models.

Problems with MLE when $\sigma_\eta / \sigma_\varepsilon$ is small:

$\Delta y_t = \Delta\beta_t + \Delta\varepsilon_t = \eta_t + \varepsilon_t - \varepsilon_{t-1} = e_t - \theta e_{t-1}$.

$\theta = \theta(\sigma_\eta/\sigma_\varepsilon)$, with $\theta(0) = 1$.

Digression: Invertibility "problem" in MA models

Suppose

$$x_t = a_t - \theta a_{t-1}$$

Then, the autocovariances of *x* are $\lambda_0 = \sigma_a^2(1+\theta^2)$ and $\lambda_1 = -\sigma_a^2\theta$

But, $\sigma_a^2(1+\theta^2) = \sigma_{\tilde{a}}^2(1+(1/\theta)^2)$ and $-\sigma_a^2\theta = -\sigma_{\tilde{a}}^2(1/\theta)$,

with $\sigma_{\tilde{a}}^2 = \sigma_a^2(1+(1/\theta)^2)^{-1}(1+\theta^2)$.

Thus, an alternative representation for *x* with exactly the same autocovariances (and Gaussian likelihood) is

$$x_t = \tilde{a}_t - (1/\theta)\tilde{a}_{t-1}.$$

These two representations are observationally equivalent. This means that the likelihoods (with σ^2 concentrated out) evaluated at θ and at $(1/\theta)$ are equal.

Returning to our problem:

$\Delta y_t = \Delta\beta_t + \Delta\varepsilon_t = \eta_t + \varepsilon_t - \varepsilon_{t-1} = e_t - \theta e_{t-1}$.

$\theta = \theta(\sigma_\eta/\sigma_\varepsilon)$, with $\theta(0) = 1$.

Because $L(\theta) = L(\theta^{-1})$, the derivative of L is zero at $\theta = 1$. Thus, the likelihood will have a local min or max at $\theta = 1$. When the true value of θ is close to 1, the global max of the likelihood is often at $\theta = 1$. Thus, $\hat{\theta}^{MLE}$ has probability mass at $\theta = 1$. In the TVP problem this translates into $\hat{\sigma}_\eta^{MLE} = 0$ with non-negligible probability if $\sigma_\eta/\sigma_\varepsilon$ is close to zero.

(Refs: Sargan and Bhargava (1983), Shephard and Harvey (1990), Shephard (1993).)

Because of problems with the MLE, it is interesting to consider other estimators.

A Median Unbiased Estimator (SW 1998). (I will talk through this estimator because it is useful in the narrow context of the TVP model, but also because the general method is used to construct confidence intervals in other many "non-standard" problems such as weak instruments, near-unit root AR models, and so forth.).

Here is the idea: Write $y_t = \beta_t + \varepsilon_t$, where $\beta_t = \beta_{t-1} + \eta_t$, and let $\alpha = (\sigma_\eta/\sigma_\varepsilon)$. We are interested in situations in which α is very close to zero. The appropriate "asymptotic nesting" is $\alpha = \gamma/T$, and we will use some asymptotic approximations where γ is held fixed as $T \rightarrow \infty$ (thus α gets closer to 0). Using this notation, we can write the model as

$$y_t = \beta_t + \varepsilon_t \quad \text{and } \beta_t = \beta_{t-1} + (\gamma/T)e_t$$

where $\text{Var}(e_t) = \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2$.

Now, think back to the TVP tests (Chow, QLR, AP) tests dicussed in lecture 2. All of the test statistics were functions of the partical sums of y. For example, in our example with $\beta_0 = 0$, the Nyblom test statistic was

$$\xi^{Nyblom} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{\sqrt{T}}\sum_{i=t}^{T} y_i)^2}{\frac{1}{T}\sum_{t=1}^{T} y_t^2}$$

and we derived the distribution of this statistic under the null that $\gamma = 0$ (β_t does not vary through time). Now, suppose we work out the distribution under the alternative (again, assuming $\beta_0 = 0$):

$$y_t = \beta_t + \varepsilon_t = (\gamma / T)\sum_{i=1}^{t} e_i + \varepsilon_t$$

so that $\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} y_t = \gamma \frac{1}{T}\sum_{t=1}^{[sT]}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{t} e_i\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} \varepsilon_t$

(repeating) $\frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} y_t = \gamma \frac{1}{T}\sum_{t=1}^{[sT]}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{t} e_i\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{[sT]} \varepsilon_t$

Now $\frac{1}{\sqrt{T}}\sum_{t=1}^{[\cdot T]} \varepsilon_t \Rightarrow \sigma_\varepsilon W(\cdot)$, and this was the term that we used when working out the null disbution.

The new term is $\gamma \frac{1}{T}\sum_{t=1}^{[\cdot T]}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{t} e_i\right) \Rightarrow B(\cdot)$, where $B(s) = \sigma_\varepsilon \gamma \int_0^s \tilde{W}(u)du$, and were W and $\tilde{W}$ are independent Wiener processes.

The key thing is that now we can derive the limiting distribution of the test statistic under both the null and the alternative. That is, as T grows large we know the distribution for $\gamma = 0$ (the null) and other values of γ as well.

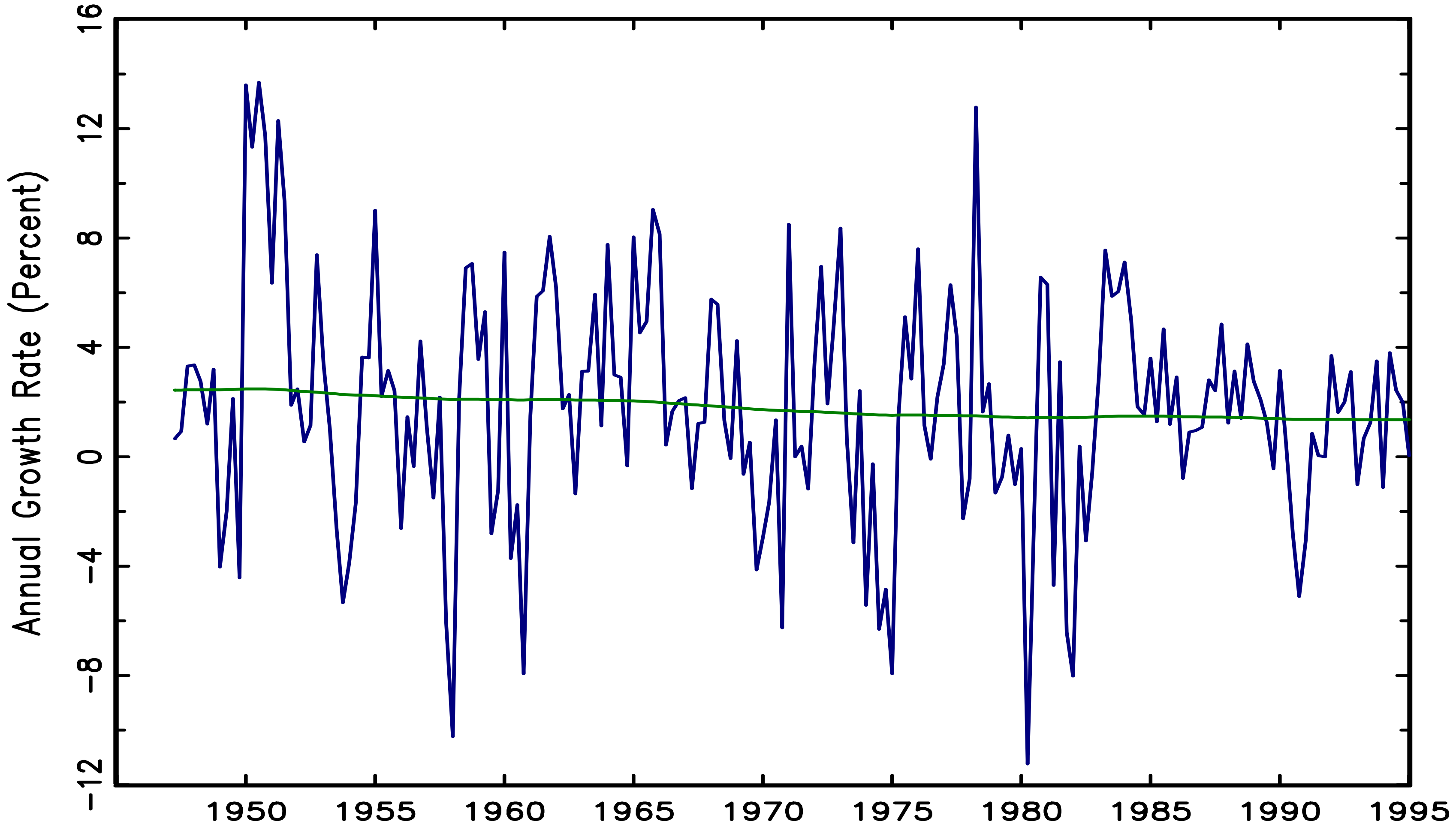
Let $A_\gamma(x)$ denote the CDF of ξ. That is $P_\gamma(\xi \leq x) = A_\gamma(x)$.

Let $\hat{\gamma}$ solve $A_{\hat{\gamma}}(\xi) = 0.5$. Then $\Pr(\hat{\gamma} \leq \gamma) = 0.5$, so that $\hat{\gamma}$ is a "median unbiased estimator" of γ.

Also $(\gamma \,|\, .95 \leq A_\gamma(\xi) \leq .05)$ is a 90% confidence interval for γ.

An example calculation (old data set):

GDP Growth Rates and estimated time varying “Mean”



$y_t = \beta_t + v_t \quad (y_t = 400\times\ln(\mathrm{GDP}_t/\mathrm{GDP}_{t-1})$

$\beta_t = \beta_{t-1} + \eta_t = \beta_{t-1} + (\gamma/T)e_t$

$v_t \sim \mathrm{AR}(1)$

$\sigma_e =$ "long-run" standard deviation of v

Data 1947 -1995

$\hat{\sigma}_\eta^{MLE} = 0$

$\hat{\gamma}^{MUB} = 4,\ \hat{\sigma}_e = 6.11\ (\hat{\sigma}_\eta = 4\times 6.11/196 = 0.12\%,\ \hat{\sigma}_{\beta_{1995-1947}} = 1.7\%)$

2 issues: Problems with MLE when $\sigma_\eta / \sigma_\varepsilon$ is small. **Computational problems in large models.**

An <u>example</u> of a Model from Lecture 11

$$Y_t = \Lambda f_t + \varepsilon_t$$

$$f_t = \phi f_{t-1} + \eta_t$$

Y_t is N×1, f_t is a scalar unobserved variable, $\Sigma_\varepsilon = \text{diag}(\sigma_i^2)$, and $\Lambda = (\lambda_1\ \lambda_2\ \ldots\ \lambda_n)'$.

Unknown Parameters: $\{\sigma_i^2\}, \{\lambda_i\}, \sigma_\eta^2, \phi$ (many if N is large).

Brute force MLE using nonlinear optimizer: Difficult
Data Augmentation-EM ("Suppose I had data on f_t) : Easy

Data Augmentation-EM

Refs: McLachlan and Krishnan (2008), Ruud (1991)

Basics:

Y: Observed data

X: Unobserved data

$f(\theta, y)$: *Y* density (or likelihood)

$f(\theta, x, y)$: Complete data density (or likelihood)

$$f(\theta, y) = \int_{x \in \boldsymbol{X}} f(\theta, x, y)dx$$

$f(x|y,\theta) = \dfrac{f(\theta, x, y)}{f(\theta, y)}$ (Conditional density of *x* given *y* evaluated at θ).

$\mathrm{L}(\theta, x, y) = \ln[f(\theta, x, y)]$ (Complete data log-likelihood)

$\mathrm{L}(\theta, y) = \ln[f(\theta, y)]$ (Incomplete data log-likelihood)

$$Q(\theta, \theta_{\mathrm{o}}, y) = \int_{x \in \boldsymbol{X}} L(\theta, x, y) f(x \mid y, \theta_o)dx = E_{\theta_o}\left\{L(\theta, x, y) \mid y\right\}$$

EM Iteration: $\theta_1 = \text{argmax}_\theta\, Q(\theta, \theta_0, y)$

Two Results:

Result 1: $\text{L}(\theta_1, y) \geq \text{L}(\theta_0, y)$

Result 2: $Q_1(\hat{\theta},\hat{\theta}, y) = 0$ if and only if $L_1(\hat{\theta}, y) = 0$, where Q_1 and L_1 are partial derivatives with respect to the first argument.

In Exponential families (normal, binomial, Bernouli, Poission, multinomial, gamma, chi-squared…), the EM iteration is easy. Let $\hat{\theta}^{MLE-CD} = h(t(X,Y))$, where $t(X,Y)$ are sufficient statistics. Then

EM Iteration: $\theta_1 = h\left(E_{\theta_0}(t(X,Y) \mid Y)\right)$

In our problem : $Y_t = \Lambda f_t + \varepsilon_t,\ f_t = \phi f_{t-1} + \eta_t$

Complete data are $\{Y_t, f_t\}$, $t = 1, \ldots, T$. The complete data Gaussian MLEs are given by the usual regression formulae:

$$\hat{\lambda}_i^{MLE-CD} = \frac{\sum_{t=1}^{T} Y_{it} f_t}{\sum_{t=1}^{T} f_t^2},\ \hat{\sigma}_i^{2,MLE-CD} = T^{-1}\left(\sum_{t=1}^{t} Y_{it}^2 - \sum_{t=1}^{t} f_t Y_{it} (\sum_{t=1}^{t} f_t^2)^{-1} \sum_{t=1}^{t} f_t Y_{it}\right)$$

$$\hat{\phi}^{MLE-CD} = \frac{\sum_{t=1}^{T} f_t f_{t-1}}{\sum_{t=1}^{T} f_{t-1}^2},\ \hat{\sigma}_\eta^{2,MLE-CD} = T^{-1}\left(\sum_{t=1}^{t} f_t^2 - \sum_{t=1}^{t} f_t f_{t-1} (\sum_{t=1}^{t} f_{t-1}^2)^{-1} \sum_{t=1}^{t} f_t f_{t-1}\right)$$

and the the expected value of these second moments conditional on the observed data, $Y_1, \ldots , Y_T$ can be computed using the Kalman Smoother.

Thus, an EM iteration is: With Λ_0, ϕ_0, $\Sigma_{\varepsilon,0}$ and $\sigma^2_{\eta,0}$

(1) Run the Kalman Smoother

(2) Compute moments as follows

(i) $E_{\theta_0}(Y_{it} f_t) = Y_{it} f_{t/T}$

(ii) $E_{\theta_0}(f_t^{\,2}) = f_{t/T}^2 + P_{t/T}$

(iii) $E_{\theta_0}(f_t f_{t-1}) = f_{t/T} f_{t-1/T} + C_{t,t-1/T}$

where $P_{t/T} = \text{var}_{\theta_0}(f_t \,|\, Y)$ and $C_{t,t-1/T} = \text{cov}_{\theta_0}(f_t f_{t-1} \,|\, Y)$, which can be computed by the Kalman Smoother

(3) Plug the results in (2) in to the usual formula for the complete data MLE to find Λ_1, ϕ_1, $\Sigma_{\varepsilon,1}$ and $\sigma^2_{\eta,1}$.

Martingale Variation: Non-Linear Models

Bag of tricks: General filtering formulae, MCMC methods for state estimation, particle filters for likelihood evaluation, Data Augmentation for simulated EM (e.g., Ruud (1991)).

Example: UC-SV

$$Y_t = \tau_t + \varepsilon_t, \qquad \tau_t = \tau_{t-1} + \eta_t$$

$$\ln(\varepsilon_t^2) = 2\ln(\sigma_{\varepsilon,t}) + \sum_{i=1}^{7} q_{\varepsilon,i,t} v_{\varepsilon,i,t}\,, \quad \ln(\eta_t^2) = 2\ln(\sigma_{\eta,t}) + \sum_{i=1}^{7} q_{\eta,i,t} v_{\eta,i,t}$$

$$\ln(\sigma_{\varepsilon,t}) = \ln(\sigma_{\varepsilon,t-1}) + \upsilon_{\varepsilon,t}, \qquad \ln(\sigma_{\eta,t}) = \ln(\sigma_{\eta,t-1}) + \upsilon_{\eta,t},$$

$$a = \left(\{\tau_t\}, \{\sigma_{\varepsilon,t}, \sigma_{\eta,t}\}, \{q_{\varepsilon,i,t}, q_{\eta,i,t}\}\right) = (a_1, a_2, a_3)$$

Martingale Variation: Non-Linear Models with “Small TVP”

Müller and Patelis (2007)

When “Nuisance” parameters are TVPs:

$\theta = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is possibly TVP.

Question: When can you ignore possible TVP in θ_2 when conducting inference about θ_1 ?

Answer: When TVP in θ_2 is sufficiently small. But how small is small? Müller and Li (2008) (general nonlinear GMM) , Li (2008) (linear model and NKPC).

Basic idea: $y_t = \alpha + x_t\beta + \varepsilon_t$

β is the parameter of interest

α is a nuissance parameter with $\alpha_t = \alpha_{t-1} + \eta_t$

Rewrite as $y_t = (\alpha_t + \bar{x}\beta) + (x_t - \bar{x})\beta + \varepsilon_t$

or $y_t = \tilde{\alpha}_t + \tilde{x}_t\beta + \varepsilon_t$

where $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} + \eta_t$

And the OLS estimator of β is $\hat{\beta} = \dfrac{\sum \tilde{x}_t y_t}{\sum \tilde{x}_t^2} = \beta + \dfrac{\sum \tilde{x}_t\tilde{\alpha}_t}{\sum \tilde{x}_t^2} + \dfrac{\sum \tilde{x}_t\varepsilon_t}{\sum \tilde{x}_t^2}$

and $$\sqrt{T}(\hat{\beta} - \beta) = \frac{\frac{1}{\sqrt{T}}\sum \tilde{x}_t\tilde{\alpha}_t}{\frac{1}{T}\sum \tilde{x}_t^2} + \frac{\frac{1}{\sqrt{T}}\sum \tilde{x}_t\varepsilon_t}{\frac{1}{T}\sum \tilde{x}_t^2}$$

The second term on the rhs is the usual source of sampling variability in the OLS estimator. Thus, the key new term is the first term on the rhs.

Term of interest: $\frac{1}{\sqrt{T}}\sum \tilde{x}_t\tilde{\alpha}_t$

Process for $\tilde{\alpha}_t$: $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} + \eta_t$

Recall MUB discussion: Power of TVP is non-trivial (not equal to 1.0) if $\sigma_\eta \sim O(T)$.

In this case $\frac{1}{\sqrt{T}}\sum \tilde{\alpha}_t \xrightarrow{d} Normal$

but (because $\tilde{x}$ has mean zero), if x process is "nice"

$$\frac{1}{\sqrt{T}}\sum \tilde{x}_t\tilde{\alpha}_t \xrightarrow{p} 0,$$

So that $\sqrt{T}(\hat{\beta}-\beta)=\dfrac{\frac{1}{\sqrt{T}}\sum \tilde{x}_t\varepsilon_t}{\frac{1}{T}\sum \tilde{x}_t^2}+o_p(1)$

Thus, if TVP in α is not so large that you would detect it with very high probability, it doesn’t matter for inference about β.