

D.C. Economists Minicourse –  
What's New in Econometrics: Time Series

Lecture 6

October 31, 2008

**The Kalman filter, Nonlinear filtering,  
and Markov Chain Monte Carlo**

# Outline

1. Models and objects of interest
2. General Formulae
3. Special Cases
4. MCMC (Gibbs)
5. Likelihood Evaluation
6. Parameter Estimation in large linear models using the EM algorithm

# 1. Models and objects of interest

General Model (Nonlinear, non-Gaussian state-space model)

(Kitagawa (1987), Fernandez-Villaverde and Rubio-Ramirez (2007))

$$y_t = H(s_t, \varepsilon_t)$$

$$s_t = F(s_{t-1}, \eta_t)$$

$\varepsilon$  and  $\eta \sim \text{iid}$

## Example 1: Linear Gaussian Model

$$y_t = Hs_t + \varepsilon_t$$

$$s_t = Fs_{t-1} + \eta_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim \text{iid}N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right)$$

## Example 2: Hamilton Regime-Switching Model

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

$$s_t = 0 \text{ or } 1 \text{ with } P(s_t = i \mid s_{t-1} = j) = p_{ij}$$

(using  $s_t = F(s_{t-1}, \eta_t)$  notation:

$$s_t = \mathbf{1}(\eta_t \leq p_{10} + (p_{11} - p_{10})s_{t-1}), \text{ where } \eta \sim U[0,1])$$

## Example 3: Stochastic volatility model

$$y_t = e^{s_t} \varepsilon_t$$

$$s_t = \mu + \phi(s_{t-1} - \mu) + \eta_t$$

## Some things you might want to calculate

Notation:  $\mathbf{Y}_t = (y_1, y_2, \dots, y_t)$ ,  $\mathbf{S}_t = (s_1, s_2, \dots, s_t)$ ,  
 $f(\cdot | \cdot)$  a generic density function.

### A. Prediction and Likelihood

(i)  $f(s_t | \mathbf{Y}_{t-1})$

(ii)  $f(y_t | \mathbf{Y}_{t-1}) \dots$  Note  $f(\mathbf{Y}_T) = \prod_{t=1}^T f(y_t | \mathbf{Y}_{t-1})$  is the likelihood

### B. Filtering: $f(s_t | \mathbf{Y}_t)$

### C. Smoothing: $f(s_t | \mathbf{Y}_T)$ .

## 2. General Formulae (Kitagawa (1987))

Model:  $y_t = H(s_t, \varepsilon_t)$ ,  $s_t = F(s_{t-1}, \eta_t)$ ,  $\varepsilon$  and  $\eta \sim \text{iid}$

A. Prediction of  $s_t$  and  $y_t$  given  $Y_{t-1}$ .

(i)

$$\begin{aligned} f(s_t | \mathbf{Y}_{t-1}) &= \int f(s_t, s_{t-1} | \mathbf{Y}_{t-1}) ds_{t-1} \\ &= \int f(s_t | s_{t-1}, \mathbf{Y}_{t-1}) f(s_{t-1} | \mathbf{Y}_{t-1}) ds_{t-1} \\ &= \int f(s_t | s_{t-1}) f(s_{t-1} | \mathbf{Y}_{t-1}) ds_{t-1} \end{aligned}$$

(ii)  $f(y_t | \mathbf{Y}_{t-1}) = \int f(y_t | s_t) f(s_t | \mathbf{Y}_{t-1}) ds_t$  (“ $t$ ” component of likelihood)

Model:  $y_t = H(s_t, \varepsilon_t)$ ,  $s_t = F(s_{t-1}, \eta_t)$ ,  $\varepsilon$  and  $\eta \sim \text{iid}$

## B. Filtering

$$f(s_t | \mathbf{Y}_t) = f(s_t | y_t, \mathbf{Y}_{t-1}) = \frac{f(y_t | s_t, \mathbf{Y}_{t-1})f(s_t | \mathbf{Y}_{t-1})}{f(y_t | \mathbf{Y}_{t-1})} = \frac{f(y_t | s_t)f(s_t | \mathbf{Y}_{t-1})}{f(y_t | \mathbf{Y}_{t-1})}$$

## C. Smoothing

$$\begin{aligned} f(s_t | \mathbf{Y}_T) &= \int f(s_t, s_{t+1} | \mathbf{Y}_T) ds_{t+1} = \int f(s_t | s_{t+1}, \mathbf{Y}_T) f(s_{t+1} | \mathbf{Y}_T) ds_{t+1} \\ &= \int f(s_t | s_{t+1}, \mathbf{Y}_t) f(s_{t+1} | \mathbf{Y}_T) ds_{t+1} = \int \left[ \frac{f(s_{t+1} | s_t) f(s_t | \mathbf{Y}_t)}{f(s_{t+1} | \mathbf{Y}_t)} \right] f(s_{t+1} | \mathbf{Y}_T) ds_{t+1} \\ &= f(s_t | \mathbf{Y}_t) \int f(s_{t+1} | s_t) \frac{f(s_{t+1} | \mathbf{Y}_T)}{f(s_{t+1} | \mathbf{Y}_t)} ds_{t+1} \end{aligned}$$

### 3. Special Cases

Model:  $y_t = H(s_t, \varepsilon_t)$ ,  $s_t = F(s_{t-1}, \eta_t)$ ,  $\varepsilon$  and  $\eta \sim \text{iid}$

General Formulae depend on  $H$ ,  $F$ , and densities of  $\varepsilon$  and  $\eta$ .

Well known special case: Linear Gaussian Model

$$y_t = Hs_t + \varepsilon_t$$

$$s_t = Fs_{t-1} + \eta_t$$

$$\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim \text{iid}N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix}\right)$$

In this case, all joint, conditional distributions and so forth are Gaussian, so that they depend only on mean and variance, and these are readily computed.

Digression: Recall that if

$$\begin{pmatrix} a \\ b \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\right),$$

then  $(a|b) \sim N(\mu_{a|b}, \Sigma_{a|b})$

where  $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(b - \mu_b)$  and  $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$ .

Interpreting  $a$  and  $b$  appropriately yields the Kalman Filter and Kalman Smoother.

(repeating) Model:  $y_t = Hs_t + \varepsilon_t$ ,  $s_t = Fs_{t-1} + \eta_t$ ,  $\begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix}\right)$

Let  $s_{t/k} = E(s_t | \mathbf{Y}_k)$ ,  $P_{t/k} = \text{Var}(s_t | \mathbf{Y}_k)$ ,  $\mu_{t/t-1} = E(y_t | \mathbf{Y}_{t-1})$ ,  $\Sigma_{t/t-1} = \text{Var}(y_t | \mathbf{Y}_{t-1})$ .

Deriving Kalman Filter:

Starting point:  $s_{t-1} | \mathbf{Y}_{t-1} \sim N(s_{t-1/t-1}, P_{t-1/t-1})$ . Then

$$\begin{pmatrix} s_t \\ y_t \end{pmatrix} | \mathbf{Y}_{t-1} \sim N\left(\begin{pmatrix} s_{t/t-1} \\ y_{t/t-1} \end{pmatrix}, \begin{pmatrix} P_{t/t-1} & P_{t/t-1}H' \\ HP_{t/t-1} & HP_{t/t-1}H' + \Sigma_\varepsilon \end{pmatrix}\right)$$

interpreting  $s_t$  as “ $a$ ” and  $y_t$  as “ $b$ ” yields the Kalman Filter.

$$\text{Model: } y_t = Hs_t + \varepsilon_t, \quad s_t = Fs_{t-1} + \eta_t, \quad \begin{pmatrix} \varepsilon_t \\ \eta_t \end{pmatrix} \sim iidN \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\varepsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right)$$

Details of KF:

- (i)  $s_{t/t-1} = Fs_{t-1/t-1}$
- (ii)  $P_{t/t-1} = FP_{t-1/t-1}F' + \Sigma_\eta$ ,
- (iii)  $\mu_{t/t-1} = Hs_{t-1/t-1}$ ,
- (iv)  $\Sigma_{t/t-1} = HP_{t-1/t-1}H' + \Sigma_\varepsilon$
- (v)  $K_t = P_{t/t-1}H'\Sigma_{t/t-1}^{-1}$
- (vi)  $s_{t/t} = s_{t-1/t-1} + K_t(y_t - \mu_{t/t-1})$
- (vii)  $P_{t/t} = (I - K_t)P_{t-1/t-1}$ .

The log-likelihood is

$$L(Y_T) = \text{constant} - 0.5 \sum_{t=1}^T \left\{ \ln |\Sigma_{t/t-1}| + (y_t - \mu_{t/t-1})' \Sigma_{t/t-1}^{-1} (y_t - \mu_{t/t-1}) \right\}$$

The Kalman Smoother (for  $s_{t/T}$  and  $P_{t/T}$ ) is derived in analogous fashion (see Anderson and Moore (2005 ), or Hamilton (1990).)

## 5. A Stochastic Volatility Model (Linear, but non-Gaussian Model) (With a slight change of notation)

$$x_t = \sigma_t e_t$$

$$\ln(\sigma_t) = \ln(\sigma_{t-1}) + \eta_t$$

or, letting  $y_t = \ln(x_t^2)$ ,  $s_t = \ln(\sigma_t)$  and  $\varepsilon_t = \ln(e_t^2)$

$$y_t = 2 s_t + \varepsilon_t$$

$$s_t = s_{t-1} + \eta_t$$

Complication:  $\varepsilon_t \sim \ln(\chi_1^2)$

## 3 ways to handle the complication

(1) Ignore it (KF is Best Linear Filter. Gaussian MLE is QMLE)  
Reference: Harvey, Ruiz, Shephard (1994)

(2) Work out analytic expressions for all the filters, etc. (Uhlig (1997) does this in a VAR model with time varying coefficients and stochastic volatility. He chooses densities and priors so that the recursive formulae yield densities and posteriors in the same family.)

(3) Numerical approximations to (2).

## Numerical Approximations: A trick and a simulation method.

Trick: Shephard (1994), Approximate the distribution of  $\varepsilon$  by a mixture of normals,  $\varepsilon_t = \sum_{i=1}^n q_{it} v_{it}$ , where  $v_{it} \sim \text{iid}N(\mu_i, \sigma_i^2)$ , and  $P(q_{it}=1)=p_i$ .

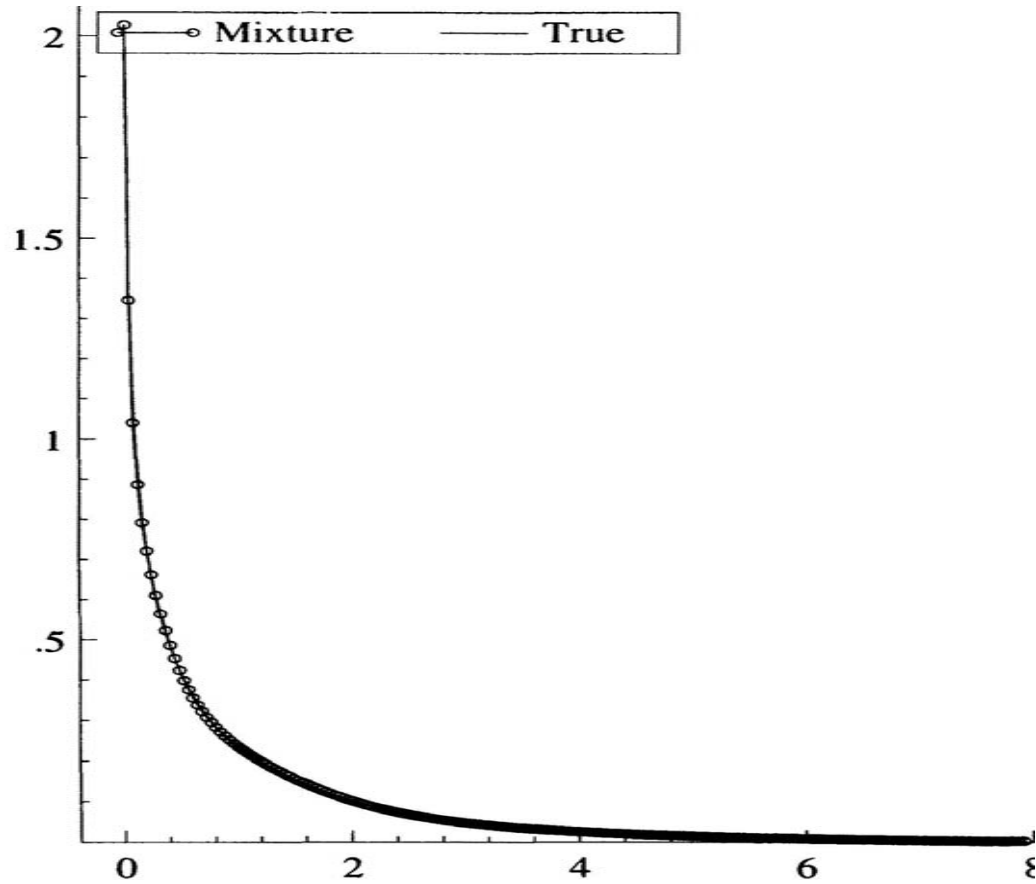
$i$	$p_i$	$\mu_i$	$\sigma_i$
1	0.00730	-10.12999	5.79596
2	0.10556	-3.97281	2.61369
3	0.00002	-8.56686	5.17950
4	0.04395	2.77786	0.16735
5	0.34001	0.61942	0.64009
6	0.24566	1.79518	0.34023
7	0.25750	-1.08819	1.26261

(numbers taken from Kim, Shephard and Chib (1998))

(Note: It seems that using only  $n=2$  does not work too poorly)

$\chi_1^2$  density and  $n=7$  mixture approximation

(picture taken from Kim, Shephard and Chib (1998))



Simulation method: MCMC methods (here Gibbs Sampling)

Some References: Casella and George (1992), Chib (2001), Geweke (2005), Koop (2003).

#### 4. Markov Chain Monte Carlo (MCMC) methods

Monte Carlo method: Let  $a$  denote a random variable with density  $f(a)$ , and suppose you want to compute  $Eg(a)$  for some function  $g$ . (Mean, standard deviation, quantile, etc.)

Suppose you can simulate from  $f(a)$ . Then  $\widehat{Eg(a)} = \frac{1}{N} \sum_{i=1}^N g(a_i)$ , where  $a_i$  are draws from  $f(a)$ . If the Monte Carlo stochastic process is sufficiently well behaved, then  $\widehat{Eg(a)} \xrightarrow{p} Eg(a)$  by the LLN.

Markov Chains: Methods for obtaining draws from  $f(a)$ . Suppose that it is difficult to draw from  $f(a)$  directly. Choose draws  $a_1, a_2, a_3, \dots$  using a Markov chain.

Draw  $a_{i+1}$  from a conditional distribution, say  $h(a_{i+1}|a_i)$ , where  $h$  has the following properties:

(1)  $f(a)$  is the invariant distribution associated with the Markov chain.  
(That is, if  $a_i$  is draw from  $f$ , then  $a_{i+1}|a_i$  is a draw from  $f$ .)

(2) Draws can't be too dependent (or else  $\widehat{Eg(a)} = \frac{1}{N} \sum_{i=1}^N g(a_i)$  will not be a good estimator of  $Eg(a)$ .)

Markov chain theory (see refs above) yields sufficient conditions on  $h$  that imply consistency and asymptotic normality of  $\widehat{Eg(a)}$ . In practice, diagnostics are used on the MC draws to see if there are problems.

How can  $h(a_{i+1}|a_i)$  be constructed so that  $f$  is invariant distribution. Gibbs sampling is one way. (Others ... )

Gibbs idea: partition  $a$  as  $a = (a^1, a^2)$ . Then  $f(a^1, a^2) = f(a^2|a^1)f(a^1)$ .

This suggests the following: given the  $i$ 'th draw of  $a$ , say  $a_i = (a_i^1, a_i^2)$ , generate  $a_{i+1}$  in two steps:

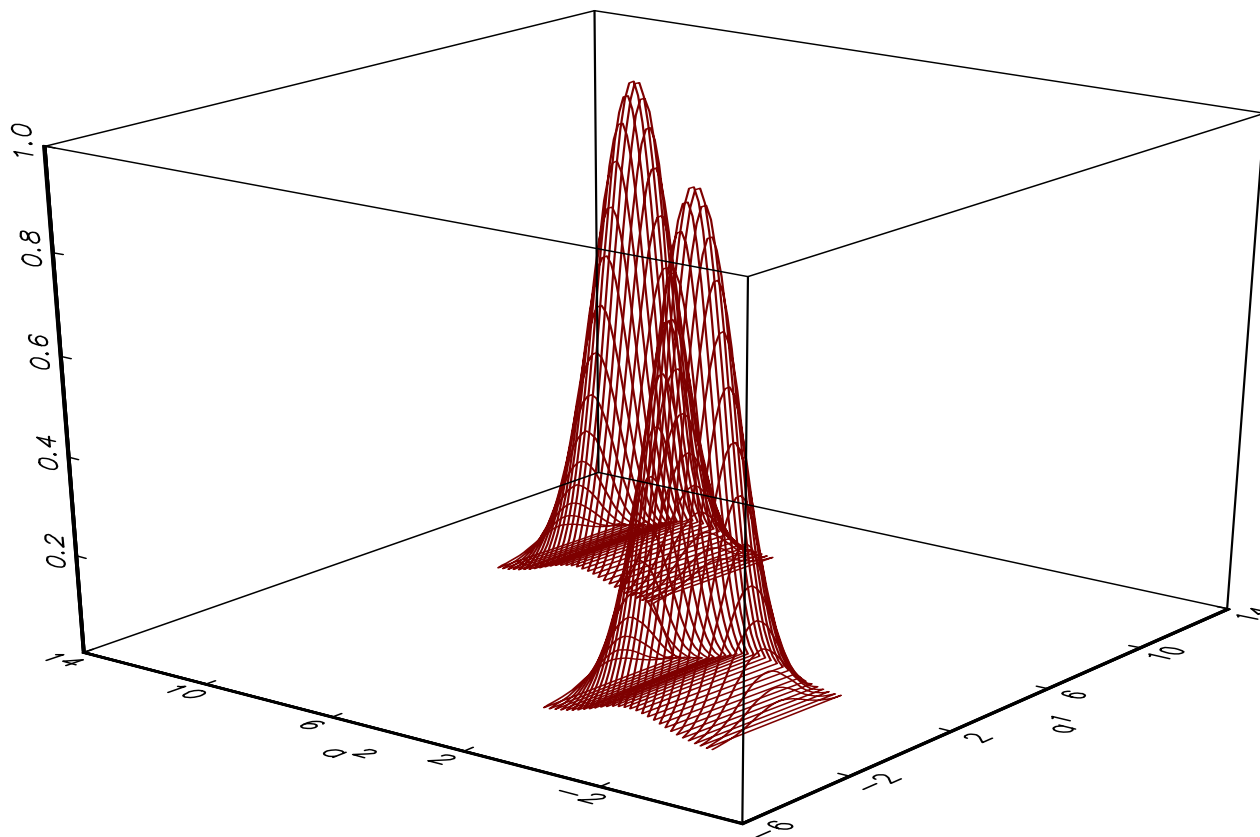
(i) draw  $a_{i+1}^1$  from  $f(a^1|a_i^2)$

(ii) draw  $a_{i+1}^2$  from  $f(a^2|a_{i+1}^1)$

Gibbs sampling is convenient when draws from  $f(a^1|a_i^2)$  and  $f(a^2|a_{i+1}^1)$  are easy.

Issues: When will this work (or when will it fail) ... draws are too correlated (requiring too many Gibbs draws for accurate Monte Carlo sample averages). Examples

(i) Bimodality:



(i) Absorbing point at  $(\tilde{a}^1, \tilde{a}^2)$ :

$$\text{Prob}(a^1 = \tilde{a}^1 \mid a_2 = \tilde{a}^2) = \text{Prob}(a_2 = \tilde{a}^2 \mid a^1 = \tilde{a}^1) = 1$$

Checking quality of approximation:  $\widehat{Eg(a)} = \frac{1}{N} \sum_{i=1}^N g(a_i)$

$$\sqrt{N}(\widehat{Eg(a)} - Eg(a)) \xrightarrow{d} N(0, V)$$

(1) 95% CI for  $Eg(a) = \widehat{Eg(a)} \pm 1.96\sqrt{\hat{V} / N}$

(2) Multiple runs from different starting values (should not differ significantly from one another)

(3) Compare  $\widehat{Eg(a)}$  based on  $N_{first}$  draws and last  $N_{last}$  draws (say first 1/3 and last 1/3 ... middle 1/3 left out). The estimates should not differ significantly from one another.

## Returning to the Stochastic Volatility Model

$$x_t = \sigma_t e_t, \quad \ln(\sigma_t) = \ln(\sigma_{t-1}) + \eta_t$$

or

$$y_t = 2 s_t + \varepsilon_t, \quad s_t = s_{t-1} + \eta_t$$

$$y_t = \ln(x_t^2), \quad \varepsilon_t = \ln(\chi_1^2) \approx \sum_{i=1}^n q_{it} v_{it}, \quad \text{where } v_{it} \sim \text{iidN}(\mu_i, \sigma_i^2), \quad \text{and } P(q_{it}=1) = p_i.$$

Smoothing Problem:  $E(\sigma_t | \mathbf{Y}_T) = E(g(s_t) | \mathbf{Y}_T)$  with  $g(s) = e^s$ :

$$\text{Let } a = \left( \{s_t\}_{t=1}^T, \{q_{it}\}_{i=1,t=1}^{7,T} \right) = (a_1, a_2)$$

Jargon: “Data Augmentation” ... add  $a_2$  to problem even though it is not of direct interest.)

Model:  $y_t = 2 s_t + \sum_{i=1}^n q_{it} v_{it}$ ,  $s_t = s_{t-1} + \eta_t$ ,  $v_{it} \sim \text{iidN}(\mu_i, \sigma_i^2)$ , and  $P(q_{it}=1)=p_i$ .

Gibbs Draws (throughout condition on  $\mathbf{Y}_T$ )

(i)  $(a_1 | a_2): \{s_t\}_{t=1}^T | \{q_{it}\}_{i=1,t=1}^{7,T}$

With  $\{q_{it}\}_{i=1,t=1}^{7,T}$  known, this is a linear Gaussian model (with known time varying “system” matrices).

$\{s_t\}_{t=1}^T | (\{q_{it}\}_{i=1,t=1}^{7,T}, Y_T)$  is normal with mean and variance easily determined by formulae analogous to Kalman-filter (see Carter, C.K. and R. Kohn (1994)).

$$(ii) (a_2 | a_1): \{q_{it}\}_{i=1,t=1}^{7,T} | \{s_t\}_{t=1}^T$$

With  $s_t$  known,  $\varepsilon_t = y_t - 2s_t$  can be calculated. So

$$\text{Prob}(q_{it} = 1 | \{s_t\}_{t=1}^T, Y_T) = \frac{f_i(\varepsilon_t)p_i}{\sum_{j=1}^7 f_j(\varepsilon_t)p_j}$$

where  $f_i$  is the  $N(\mu_i, \sigma_i^2)$  density.

## More Complicated Examples:

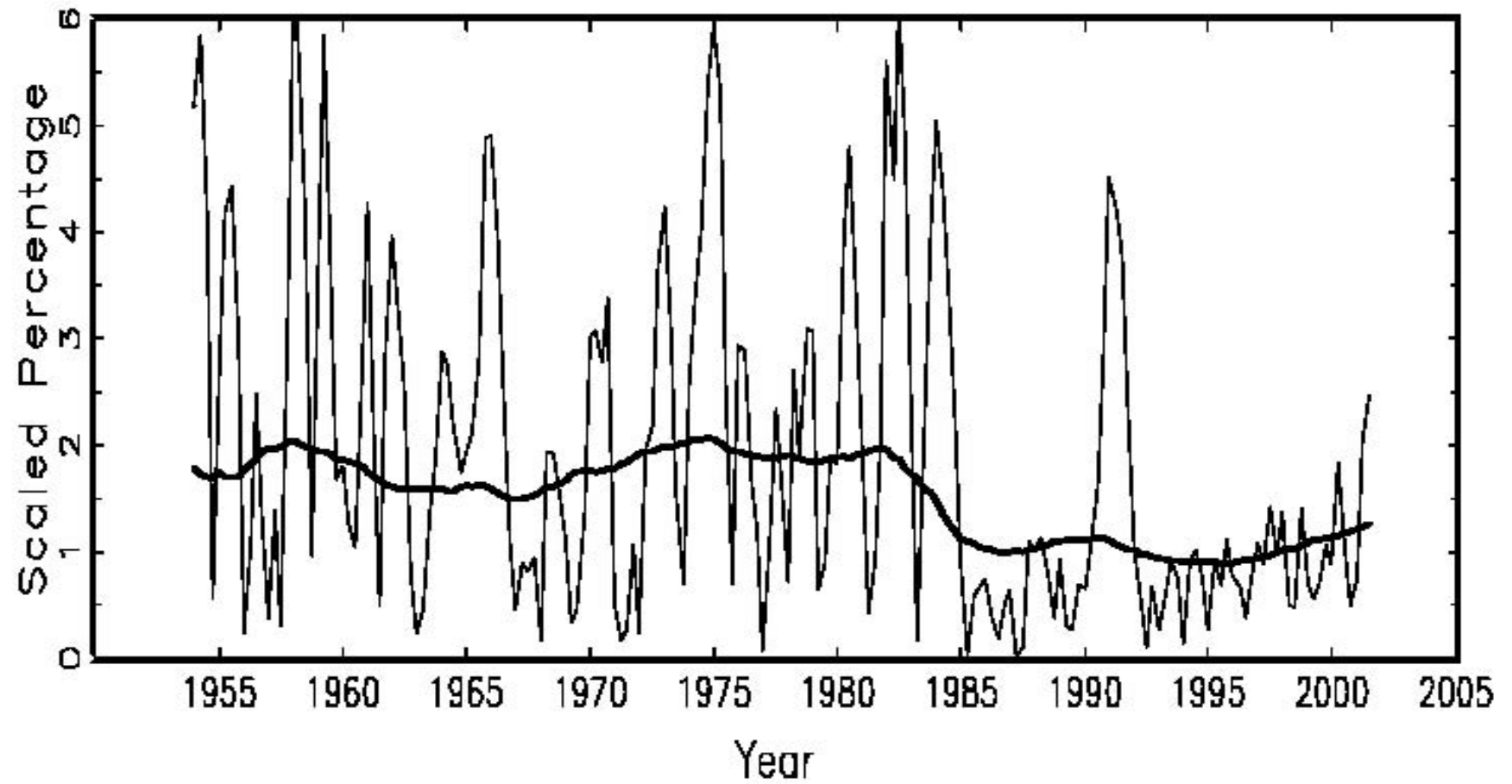
$$\text{TVP-VAR-SV Model: } y_t = \sum_{i=1}^p \Phi_t y_{t-i} + e_t \quad (e_t \sim \text{SV})$$

(VAR) Cogley and Sargent (2005), Uhlig (1997), (SVAR) Primiceri (2005), (Markov Switching VAR) Sims and Zha (2006).

Simple univariate version: SW (2002) –  $y_t$  is quarterly GDP growth rates.

Compute model's implied SD of  $y_t + y_{t-1} + y_{t-2} + y_{t-3} = \text{Annual growth rate}$ .

## A. GDP



UC-SV:  $Y_t = \tau_t + \varepsilon_t$ ,  $\tau_t = \tau_{t-1} + \eta_t$  ( $\varepsilon_t$  and  $\eta_t \sim \text{SV}$ )

Stock and Watson (2007),

Note:  $\Delta Y_t = \eta_t + \varepsilon_t - \varepsilon_{t-1}$ , so with constant volatility  $Y_t \sim \text{IMA}(1,1)$ , and SV yields a time varying MA coefficient.

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

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

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

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

Gibbs Draws:

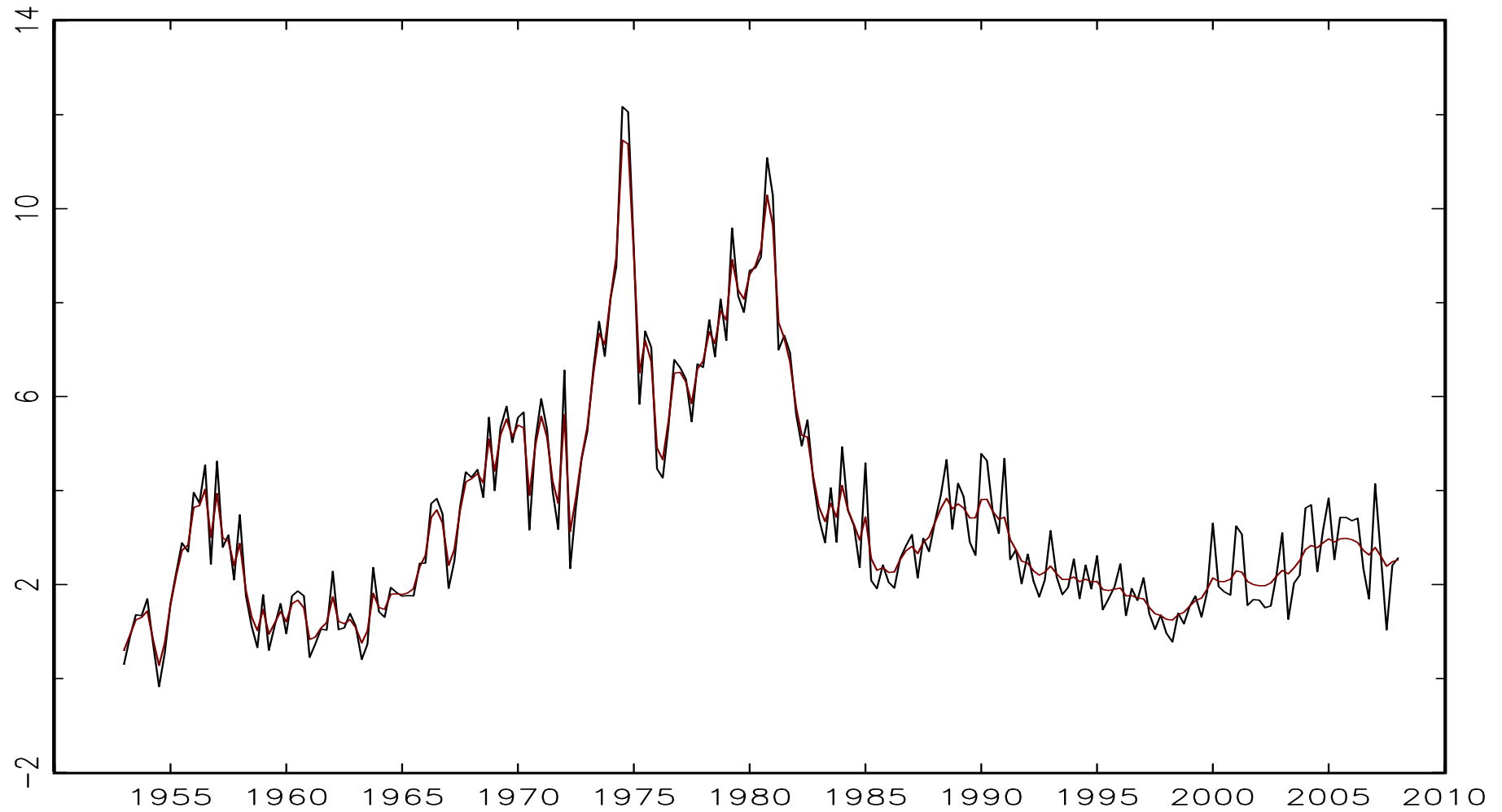
$\{ \tau_t \} \mid \{ \sigma_{\varepsilon,t}, \sigma_{\eta,t} \}, \{ q_{\varepsilon,i,t}, q_{\eta,i,t} \}, Y_T$ : “Kalman filter” – UC Model

$\{ \sigma_{\varepsilon,t}, \sigma_{\eta,t} \} \mid \{ \tau_t \}, \{ q_{\varepsilon,i,t}, q_{\eta,i,t} \}, Y_T$ : “Kalman filter” – SV (as above)

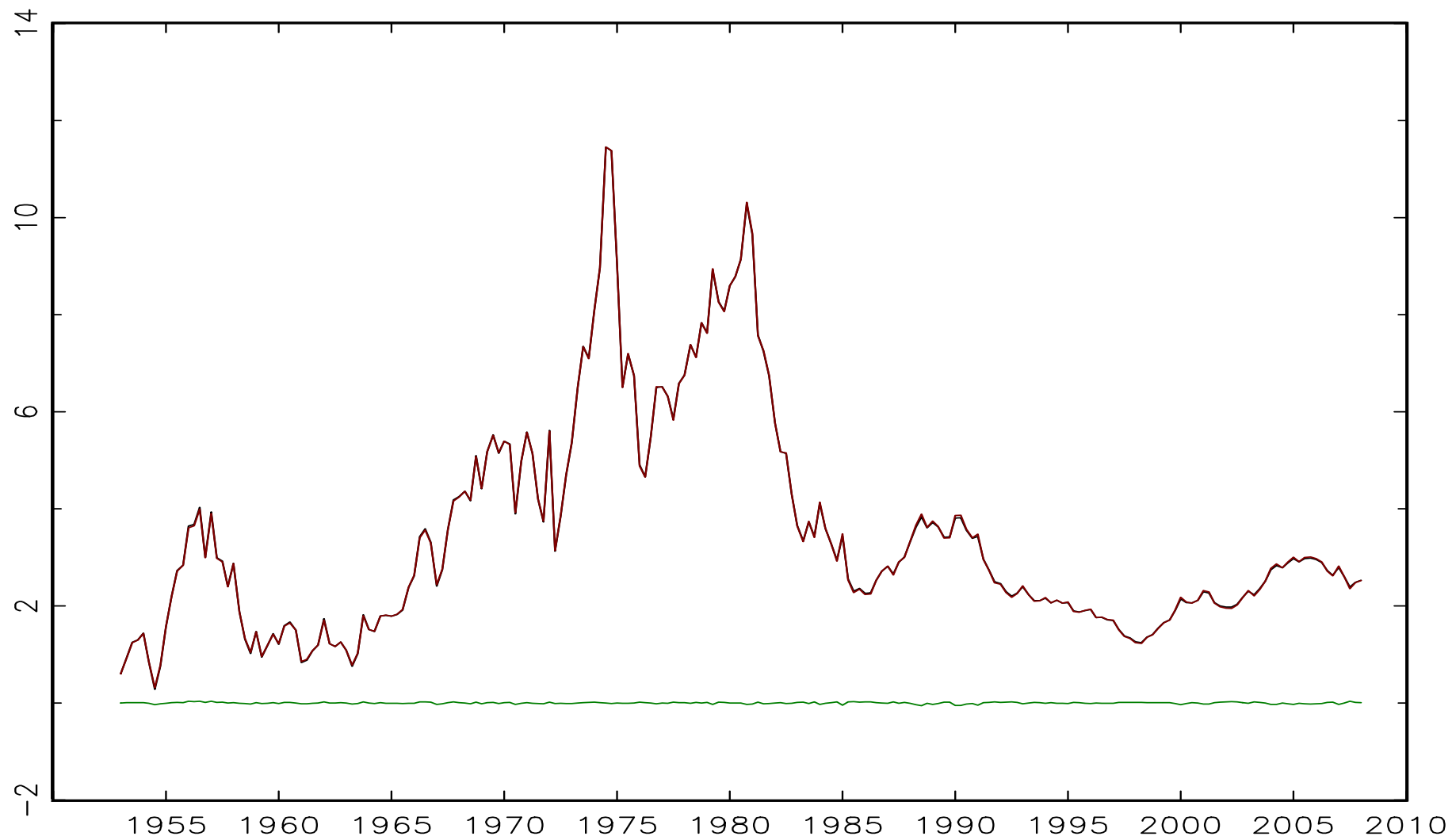
$\{ q_{\varepsilon,i,t}, q_{\eta,i,t} \} \mid \{ \tau_t \}, \{ \sigma_{\varepsilon,t}, \sigma_{\eta,t} \}, Y_T$ : Mixture indicator draws ... as above

# Inflation (GDP Deflator) and smoothed estimate of $\tau$

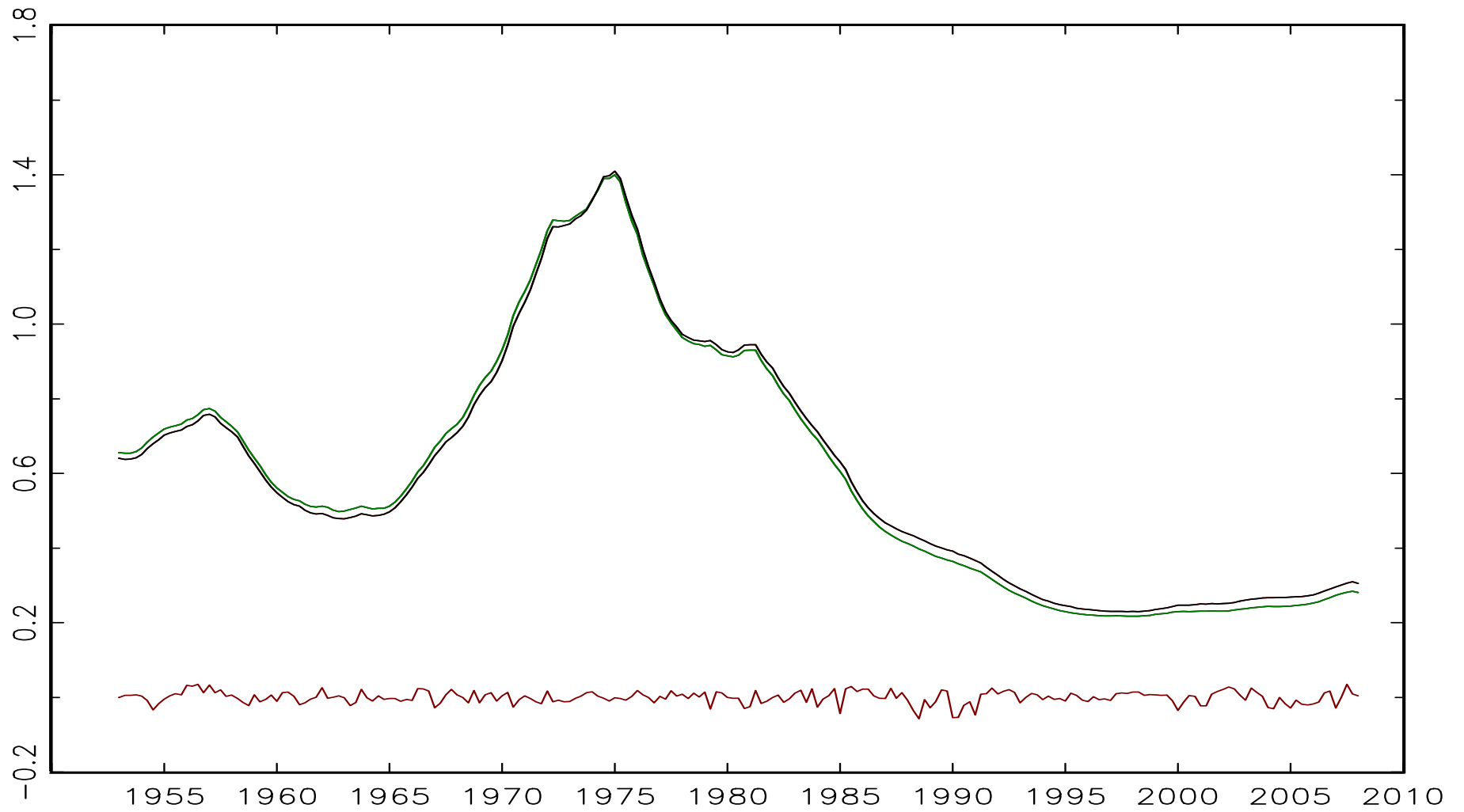
( $N = 10,000$ , burnin = 1000)



# Estimates of $\tau$ from two independent sets of draws



# Estimates of $\sigma_\eta$ from two independent sets of draws



$$\widehat{Eg(a)} = \frac{1}{N} \sum_{i=1}^N g(a_i); \quad \sqrt{N}(\widehat{Eg(a)} - Eg(a)) \xrightarrow{d} N(0, V)$$

Average values over all dates

	<i>Serial Correlation in <math>g(a_i)</math></i>	$\sqrt{V / N}$	$\frac{\sqrt{V / n}}{\widehat{Eg(a)}}$
$\tau$	0.19	0.025	0.7%
$\sigma_\eta$	0.57	0.018	3%

What can go wrong (2): “Absorbing Barrier” (or “just getting stuck”)

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

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

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

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

Cecchetti, Hooper, Kasman, Schoenholtz and Watson (2007)

What happens if  $\sigma_{\eta,t}$  gets very small?

## Computing the likelihood: Particle filtering

Model:  $y_t = H(s_t, \varepsilon_t)$ ,  $s_t = F(s_{t-1}, \eta_t)$ ,  $\varepsilon$  and  $\eta \sim \text{iid}$

The “ $t$ ’th component” of likelihood:  $f(y_t | \mathbf{Y}_{t-1}) = \int f(y_t | s_t) f(s_t | \mathbf{Y}_{t-1}) ds_t$

Often  $f(y_t | s_t)$  is known, and the challenge is  $f(s_t | Y_{t-1})$ . Particle filters use simulation methods to draw samples from  $f(s_t | Y_{t-1})$ , say  $(s_{1t}, s_{2t}, \dots, s_{nt})$ , where  $s_{it}$  is called a “particle.” The  $t$ ’th component of the likelihood can

then be approximated as  $\widehat{f(y_t | Y^{t-1})} = \frac{1}{n} \sum_{i=1}^n f(y_t | s_{it})$ .

Methods for computing draws utilize the structure of the particular problem under study. Useful references include Kim, Shephard and Chib (1998), Chib, Nardari and Shephard (2002), Pitt and Shephard (1999), and Fernandez-Villaverde and Rubio-Ramirez (2007).

## 6. Parameter Estimation in large linear models using the EM algorithm

An example from Lecture 11

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

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

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

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)

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

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

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

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

$$Q(\theta, \theta_0, y) = \int_{x \in X} L(\theta, x, y) f(x | y, \theta_0) dx = E_{\theta_0} \{L(\theta, x, y) | y\}$$

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

Two Results:

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

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

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

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

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

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

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

Thus, an EM iteration is: With  $\Lambda_0$ ,  $\phi_0$ ,  $\Sigma_{\varepsilon,0}$  and  $\sigma_{\eta,0}^2$

(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  $\Lambda_1$ ,  $\phi_1$ ,  $\Sigma_{\varepsilon,1}$  and  $\sigma_{\eta,1}^2$ .

# Outline

1. Models and objects of interest
2. General Formulae
3. Special Cases
4. MCMC (Gibbs)
5. Likelihood Evaluation
6. Parameter Estimation in large linear models using the EM algorithm