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
1. Models and objects of interest

General Model (Nonlinear, non-Gaussian state-space model) 
(Kitagawa (1987), Fernandez-Villaverde and Rubio-Ramirez (2007))

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

Example 1: Linear Gaussian Model

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

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

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: \( Y_t = (y_1, y_2, \ldots, y_t) \), \( S_t = (s_1, s_2, \ldots, s_t) \),
\( f(\cdot | \cdot) \) a generic density function.

A. Prediction and Likelihood

(i) \( f(s_t | Y_{t-1}) \)

(ii) \( f(y_t | Y_{t-1}) \) … Note \( f(Y_T) = \prod_{t=1}^{T} f(y_t | Y_{t-1}) \) is the likelihood

B. Filtering: \( f(s_t | Y_t) \)

C. Smoothing: \( f(s_t | Y_T) \).
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)  
\[
 f(s_t \mid Y_{t-1}) = \int f(s_t, s_{t-1} \mid Y_{t-1}) ds_{t-1} \\
 = \int f(s_t \mid s_{t-1}, Y_{t-1}) f(s_{t-1} \mid Y_{t-1}) ds_{t-1} \\
 = \int f(s_t \mid s_{t-1}) f(s_{t-1} \mid Y_{t-1}) ds_{t-1}
\]

(ii)  
\[
 f(y_t \mid Y_{t-1}) = \int f(y_t \mid s_t) f(s_t \mid Y_{t-1}) ds_t \quad ("t" \ \text{component of likelihood})
\]
Model: $y_t = H(s_t, \epsilon_t)$, $s_t = F(s_{t-1}, \eta_t)$, $\epsilon$ and $\eta$ ~ iid

B. Filtering

$$f(s_t | Y_t) = f(s_t | y_t, Y_{t-1}) = \frac{f(y_t | s_t, Y_{t-1}) f(s_t | Y_{t-1})}{f(y_t | Y_{t-1})} = \frac{f(y_t | s_t) f(s_t | Y_{t-1})}{f(y_t | Y_{t-1})}$$

C. Smoothing

$$f(s_t | Y_T) = \int f(s_t, s_{t+1} | Y_T) ds_{t+1} = \int f(s_t | s_{t+1}, Y_T) f(s_{t+1} | Y_T) ds_{t+1}$$

$$= \int f(s_t | s_{t+1}, Y_T) f(s_{t+1} | Y_T) ds_{t+1} = \int \left[ \frac{f(s_{t+1} | s_t) f(s_t | Y_t)}{f(s_{t+1} | Y_T)} \right] f(s_{t+1} | Y_T) ds_{t+1}$$

$$= f(s_t | Y_t) \int f(s_{t+1} | s_t) \frac{f(s_{t+1} | Y_T)}{f(s_{t+1} | Y_t)} ds_{t+1}$$
3. Special Cases

Model: \( y_t = H(s_t, \epsilon_t), \quad s_t = F(s_{t-1}, \eta_t), \quad \epsilon \) and \( \eta \sim \text{iid} \)

General Formulae depend on \( H, F, \) and densities of \( \epsilon \) and \( \eta \).

Well known special case: Linear Gaussian Model

\[
y_t = Hs_t + \epsilon_t \\
s_t = Fs_{t-1} + \eta_t \\
\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \sim \text{iid} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\eta \end{pmatrix} \right)
\]

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

Let \( s_{t/k} = E(s_t | Y_k), \quad P_{t/k} = \text{Var}(s_t | Y_k), \quad \mu_{t/t-1} = E(y_t | Y_{t-1}), \quad \Sigma_{t/t-1} = \text{Var}(y_t | Y_{t-1}). \)

Deriving Kalman Filter:
Starting point: \( s_{t-1} | Y_{t-1} \sim N(s_{t-1/t-1}, P_{t-1/t-1}). \) Then

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

interpreting \( s_t \) as “a” and \( y_t \) as “b” yields the Kalman Filter.
Model: \( y_t = Hs_t + \varepsilon_t, \quad s_t = Fs_{t-1} + \eta_t, \quad \left( \begin{array}{c} \varepsilon_t \\ \eta_t \end{array} \right) \sim iidN \left( \begin{array}{c} 0 \\ 0 \end{array}, \begin{pmatrix} \Sigma_{\varepsilon} & 0 \\ 0 & \Sigma_{\eta} \end{pmatrix} \right) \)

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) \( \mu_{t/t-1} = Hs_{t/t-1} \),
(iv) \( \Sigma_{t/t-1} = HP_{t/t-1} H' + \Sigma_{\varepsilon} \)
(v) \( K_t = P_{t/t-1} H' \Sigma_{t/t-1}^{-1} \)
(vi) \( s_{t/t} = s_{t/t-1} + K_t (y_t - \mu_{t/t-1}) \)
(vii) \( P_{t/t} = (I - K_t) P_{t/t-1} \).
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 $P(q_{it}=1)=p_i$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$p_i$</th>
<th>$\mu_i$</th>
<th>$\sigma_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00730</td>
<td>-10.12999</td>
<td>5.79596</td>
</tr>
<tr>
<td>2</td>
<td>0.10556</td>
<td>-3.97281</td>
<td>2.61369</td>
</tr>
<tr>
<td>3</td>
<td>0.00002</td>
<td>-8.56686</td>
<td>5.17950</td>
</tr>
<tr>
<td>4</td>
<td>0.04395</td>
<td>2.77786</td>
<td>0.16735</td>
</tr>
<tr>
<td>5</td>
<td>0.34001</td>
<td>0.61942</td>
<td>0.64009</td>
</tr>
<tr>
<td>6</td>
<td>0.24566</td>
<td>1.79518</td>
<td>0.34023</td>
</tr>
<tr>
<td>7</td>
<td>0.25750</td>
<td>-1.08819</td>
<td>1.26261</td>
</tr>
</tbody>
</table>

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

(Note: It seems that using only $n=2$ does not work too poorly)
$\chi^2$ density and $n=7$ mixture approximation

(picture taken from Kim, Shephard and Chib (1998))
Simulation method: MCMC methods (here Gibbs Sampling)


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 $\hat{Eg(a)} = \frac{1}{N} \sum_{i=1}^{N} g(a_i)$, where $a_i$ are draws from $f(a)$. If the Monte Carlo stochastic process is sufficiently well behaved, then $\frac{\hat{Eg(a)}}{p} \rightarrow Eg(a)$ 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 $a_1, a_2, a_3, \ldots$ 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 $\frac{1}{N} \sum_{i=1}^{N} g(a_i)$ will not be a good estimator of $\widehat{Eg}(a)$.)

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 \( f(a_1^1|a_i^2) \)

(ii) draw \( a_{i+1}^2 \) from \( f(a_2^2|a_{i+1}^1) \)

Gibbs sampling is convenient when draws from \( f(a_1^1|a_i^2) \) and \( f(a_2^2|a_{i+1}^1) \) 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: $\hat{Eg}(a) = \frac{1}{N} \sum_{i=1}^{N} g(a_i)$

$\sqrt{N} (\hat{Eg}(a) - Eg(a)) \overset{d}{\rightarrow} N(0, V)$

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

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

(3) Compare $\hat{Eg}(a)$ 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

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

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

Smoothing Problem: \( E(\sigma_t \mid Y_T) = E(g(s_t) \mid Y_T) \) with \( g(s) = e^s \):

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

Jargon: “Data Augmentation” … add \( a_2 \) to problem even though it is not of direct interest.)
Model: \( y_t = 2s_t + \sum_{i=1}^{n} q_{it}v_{it} \), \( s_t = s_{t-1} + \eta_t \), \( v_{it} \sim \text{iidN}(\mu_i, \sigma_i^2) \), and \( P(q_{it}=1)=p_i \).

Gibbs Draws (throughout condition on \( Y_T \))

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

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

\( \{s_t\}_{t=1}^{T} \mid (\{q_{it}\}_{i=1,t=1}^{7,T}, Y_T) \) is normal with mean and variance easily determined by formulae analogous to Kalman-filter (see Carter, C.K. and R. Kohn (1994)).
(ii) \((a_2 \mid a_1)\): \(\{q_{it}\}_{i=1,t=1}^{7,T} \mid \{s_t\}_{t=1}^T\)

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

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

where \(f_i\) is the \(\text{N}(\mu_i, \sigma_i^2)\) density.
More Complicated Examples:

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


Compute model’s implied SD of \( y_t + y_{t-1} + y_{t-2} + y_{t-3} = \) Annual growth rate.
A. GDP

Scaled Percentage

Year


Lecture 5 - 27, July 21, 2008
UC-SV: $Y_t = \tau_t + \epsilon_t, \quad \tau_t = \tau_{t-1} + \eta_t \quad (\epsilon_t \text{ and } \eta_t \sim SV)$

Stock and Watson (2007),

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

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

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

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

Gibbs Draws:

\{ \tau_t \} | \{ \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} \} | \{ \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} \} | \{ \tau_t \}, \{ \sigma_{\varepsilon,t}, \sigma_{\eta,t} \}, Y_T: Mixture indicator draws … as above
Inflation (GDP Deflator) and smoothed estimate of $\tau$

$(N = 10,000, \text{burnin} = 1000)$
Estimates of $\tau$ from two independent sets of draws
Estimates of $\sigma_\eta$ from two independent sets of draws
\[ \overline{Eg(a)} = \frac{1}{N} \sum_{i=1}^{N} g(a_i) ; \quad \sqrt{N} (\overline{Eg(a)} - Eg(a)) \xrightarrow{d} N(0,V) \]

Average values over all dates

<table>
<thead>
<tr>
<th>Serial Correlation in ( g(a_i) )</th>
<th>( \sqrt{V / N} )</th>
<th>( \sqrt{V / n} \overline{Eg(a)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>0.19</td>
<td>0.025</td>
</tr>
<tr>
<td>( \sigma_\eta )</td>
<td>0.57</td>
<td>0.018</td>
</tr>
</tbody>
</table>
What can go wrong (2): “Absorbing Barrier” (or “just getting stuck”)

\[ Y_t = \tau_t + \varepsilon_t, \]
\[ \tau_t = \tau_{t-1} + \eta_t \]

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

\[ \ln(\sigma_{\varepsilon,t}) = \ln(\sigma_{\varepsilon,t-1}) + v_{\varepsilon,t}, \]
\[ \ln(\sigma_{\eta,t}) = \ln(\sigma_{\eta,t-1}) + v_{\eta,t}, \]

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

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: \( f(y_t \mid Y_{t-1}) = \int f(y_t \mid s_t) f(s_t \mid Y_{t-1}) ds_t \)

Often \( f(y_t \mid s_t) \) is known, and the challenge is \( f(s_t \mid Y_{t-1}) \). Particle filters use simulation methods to draw samples from \( f(s_t \mid Y_{t-1}) \), say \( (s_{1t}, s_{2t}, \ldots s_{nt}) \), where \( s_{it} \) is called a “particle.” The \( t’ \)th component of the likelihood can then be approximated as \( \frac{1}{n} \sum_{i=1}^{n} f(y_t \mid s_{it}) \).