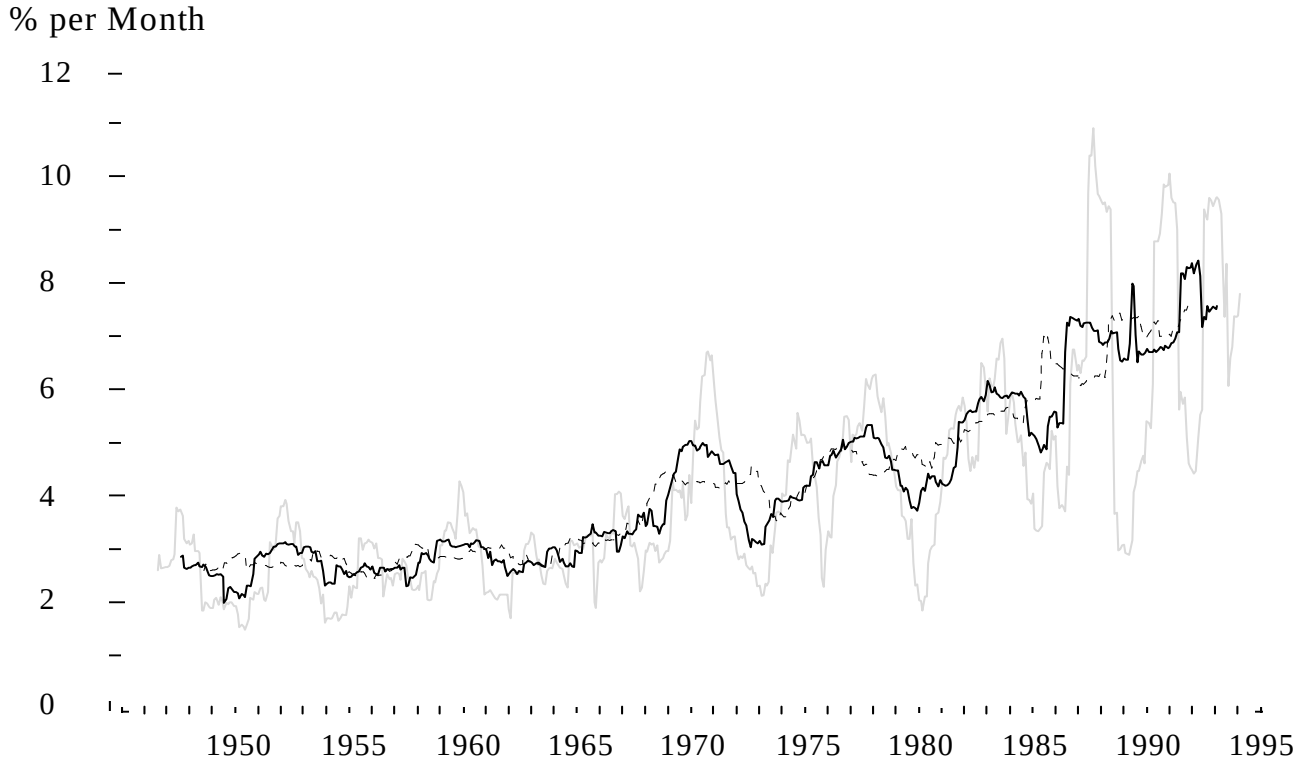
Hamilton/Lindgren Regime Switching Models

Idea: The parameters of a stochastic process shift abruptly when an underlying state variable shifts. Example: 1. Recession-recovery represents the two states. In recession expected growth is higher. 2. High and Low volatility on stock market. These states are exogenous to the stock market and represent shifts in the flow of information. (The exogeneity can be relaxed). Expected volatility is higher in one of the states.



Markov Chains

Let s_t be a stochastic variable that can take on integer values in the finite set $\Omega = \{0,1,\ldots,N\}$. Furthermore, assume that the probability

seen from time *t*-1 that $s_t = i \in \Omega$ is determined *only* by the current value of s_{t-1}. So $P\{s_t = i | s_{t-1} = j\} = P\{s_t = i | s_{t-1} = j, \ldots, s_{t-s}, \Theta_{t-1}\}$. Call the probability $P\{s_t = i | s_{t-1} = j\} \equiv p_{ji}$. Such (state) variable is defined as a (first order) *Markov chain*.

Transition matrix

Assume that we know the probability $P\{s_{t-1} = i | \Theta_{t-1}\}$ for all *i*. Then clearly

$$P\{s_t = j | \Theta_{t-1}\} = p_{1j} P\{s_{t-1} = 1 | \Theta_{t-1}\} + \ldots + p_{Nj} P\{s_{t-1} = N | \Theta_{t-1}\} \quad (1)$$

If we collect the probabilities p_{ij} in a matrix we can define the N x N matrix

$$\mathbf{P} \equiv \begin{bmatrix} p_{11} & \cdots & p_{N1} \\ \vdots & \ddots & \vdots \\ P_{1N} & \cdots & P_{NN} \end{bmatrix} \quad (2)$$

which is called the *transition matrix.* Note that the columns but not the rows have to sum to unity. Now stack the probabilities $P\{s_{t-1} = i | \Theta_{t-1}\}$ in a N x 1 vector defined as $\xi_{t-1|t-1}$ as we can write in vector notation

$$\xi_{t+1|t} = \mathbf{P}\xi_{t|t} \quad (3)$$

and

$$\xi_{t+m|t} = \mathbf{P}^m \xi_{t|t} \quad (4)$$

Is the first order assumption restrictive?

Above we assumed that the probability that the state variable takes some value *i* next period only depends on the current state. This is not

restrictive since we can redefine the states so that this is satisfied. For example, assume that s_t is 0 or 1 and that

$$P\{s_{t+1}=1|\Theta_t\}=\begin{cases}p_{111} & if\ s_{t.}=1, s_{t-1}=1\\ p_{011} & if\ s_{t.}=1, s_{t-1}=0\\ p_{101} & if\ s_{t.}=0, s_{t-1}=1\\ p_{001} & if\ s_{t.}=0, s_{t-1}=0\end{cases} \tag{5}$$

This definition of the states violates that (first order) Markov assumption. But now consider the following redefinition of the state variable

$$ss_t \equiv \begin{cases}0 & if\ s_{t.}=1, s_{t-1}=1\\ 1 & if\ s_{t.}=1, s_{t-1}=0\\ 2 & if\ s_{t.}=0, s_{t-1}=1\\ 3 & if\ s_{t.}=0, s_{t-1}=0\end{cases} \tag{6}$$

We can now write the transition matrix for *ss*

$$\mathbf{P}=\begin{bmatrix} p_{111} & p_{011} & 0 & 0\\ 0 & 0 & p_{101} & p_{001}\\ 1-p_{111} & 1-p_{011} & 0 & 0\\ 0 & 0 & 1-p_{101} & 1-p_{001}\end{bmatrix} \tag{7}$$

This "super state" variable satisfies the Markov conditions.

Absorbing state

If we by relabeling different states can write $\mathbf{P}$ as

$$\mathbf{P}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{0} & \mathbf{C}\end{bmatrix} \tag{8}$$

where **A** is a K x K matrix and **0** is a (N-K x K) matrix. Then if the state variable takes any of the first *k* states it can never return to states K+1 to N. The first states then an absorbing set of states and **P** is *reducible.*

Ergodicity

To call the Markov chain to be ergodic we want the importance of the current state for future state probabilities to die away as the horizon goes to infinity. Using (4), we want

$$\xi = \lim_{T \to \infty} \mathbf{P}^T \xi_{t|t} \tag{9}$$

to be the same for any $\xi_{t|t}$. This limit is then the unconditional expected state probabilities, or the *ergodic* state probabilities. This vector satisfies

$$\xi = \mathbf{P}\xi \tag{10}$$

which we understand from (9). (10) together with the obvious requirement **1'**ξ=1 gives us sufficient information to solve for the unconditional state probabilities.

A sufficient condition for **P** to produce an ergodic Markov chain is that one eigenvalue is unity and the others are inside the unit circle.

The Simplest Regime Switching Model

Now let us consider a stochastic process y_t which is a random walk with a drift term that depends on an unobservable state s_t which is zero or unity.

$$y_{t+1} = \alpha_0 + \alpha_1 s_{t+1} + y_t + \varepsilon_{t+1} \tag{11}$$

with ε is i.i.d. $N(0,\sigma^2)$ sequence. The state has a transition matrix

$$\mathbf{P} = \begin{bmatrix} p & 1-q \\ 1-p & q \end{bmatrix} \tag{12}$$

The Recursive Hamilton Filter

Assume we have a sequence of observations on y_t for $t \in \{0,T)$. We now want to calculate the likelihood of these observations as a function of the parameter vector $\theta = \{\alpha_0, \alpha_1, \sigma^2, p, q,\}$ and the sequence of state probabilities $P\{s_t = 0 | y_t, \ldots, y_0; \theta, P(s_{t-1} = 1)\}$ for $t=0,\ldots T$.

First denote the vector $\{P(s_t = 0 | y_t, \ldots, y_0), P(s_t = 1 | y_t, \ldots, y_0)\}$ by $P(s_t | y_t)$ and let $P(s_{t+1}, s_t | y_t)$ denote the 4 x 1 vector $\{P(s_{t+1} = 0, s_t = 0 | y_t, \ldots, y_0), P(s_{t+1} = 1, s_t = 0 | y_t, \ldots, y_0), \ldots\}$

For the purpose we construct a filter takes as

input $P(s_{t-1} | y_{t-1})$ (13)

and produces as

output $P(s_t | y_t)$ and $f(y_t | y_{t-1}, \ldots, y_0)$. (14)

The filter is constructed in five steps.

Step 1

Calculate $P(s_t, s_{t-1} | y_{t-1})$. This is done by using the rule P(A and B)= P(B|A)P(A). So in this case,

$$P\left(s_t = i, s_{t-1} = j \middle| y_{t-1}\right) = P\left(s_t = i \middle| s_{t-1} = j\right) P\left(s_{t-1} = j \middle| y_{t-1}\right) \tag{15}$$

Gauss programming suggestion. Element by element multiplication of $P\left(s_t, s_{t-1} \middle| y_{t-1}\right) = \mathbf{P}.*P\left(s_{t-1} \middle| y_{t-1}\right)$ produces a 2 x 2 matrix with the 4 elements of (15). This matrix gives the probabilities of s_t , s_{t-1} equalling

$$\begin{bmatrix} 0,0 & 1,0 \\ 0,1 & 1,1 \end{bmatrix} \tag{16}$$

Step 2

Calculate the joint density $f\left(y_t, s_t, s_{t-1} \middle| y_{t-1}\right)$ for the four possible combinations of s_t and s_{t-1}.

$$f\left(y_t, s_t = i, s_{t-1} = j \middle| y_{t-1}\right) = f\left(y_t \middle| s_t = i, s_{t-1} = j, y_{t-1}\right) P\left(s_t = i, s_{t-1} = j \middle| y_{t-1}\right) \tag{17}$$

Gauss programming hint. Create a 2 x 2 matrix of expected values of y_t for the state sequences in (16). I.e.,

$$M_t = \begin{bmatrix} \alpha_0 + y_{t-1} & \alpha_0 + \alpha_1 + y_{t-1} \\ \alpha_0 + y_{t-1} & \alpha_0 + \alpha_1 + y_{t-1} \end{bmatrix} \tag{18}$$

then (2 pi sigma^2)^-.5 *exp(-(1/(2*sigma^2)*(M_t-y_t)) produces the 2 x 2 matrix of $f\left(y_t \middle| s_t, s_{t-1}, y_{t-1}\right)$ for the histories in (16).

Step 3

Calculate the density $f\left(y_t \middle| y_{t-1}\right)$ by summing all elements of $f\left(y_t, s_t, s_{t-1} \middle| y_{t-1}\right)$.

Gauss programming hint. Sumc(sumc($f\left(y_t, s_t, s_{t-1} \middle| y_{t-1}\right)$)). Make sure this value is stored.

Step 4

Now we want to calculate $P\left(s_t, s_{t-1} \middle| y_t\right)$. For this we use Bayes law

P(A|B)=P(AB)/P(B). So in this case

$$P\left(s_t = i, s_{t-1} = j \middle| y_t\right) = \frac{f\left(y_t, s_t = i, s_{t-1} = j \middle| y_{t-1}\right)}{f\left(y_t \middle| y_{t-1}\right)} \tag{19}$$

Gauss programming hint. Divide the matrix that came from step 2 by the number that resulted from step 3.

Step 5

By summing $P\left(s_t, s_{t-1} \middle| y_t\right)$ over s_{t-1} we produce $P\left(s_t \middle| y_t\right)$ which is the output of the filter.

Gauss programming hint. Sumc($P\left(s_t, s_{t-1} \middle| y_t\right)$).

By repeating this until from t=1 until T we get a sequence of state probabilities and a sequence of likelihoods. $f(y_t | y_{t-1}, \ldots y_0)$. The only problem is that we need a starting value for the filter i.e., $P(s_0 | y_0)$. The simplest approach is here to use the unconditional (ergodic) probability which is determined by **P.**

Estimation

The parameters of the model is estimated by maximising the log likelihood function. Note also that we may use the EM algorithm, which is convenient in some cases.

Newton-Raphson

Suppose we are looking for an optimum of the function $f(\mathbf{x}), \mathbf{x} \in R^n$. A standard way to do this numerically is to apply the *Newton-Raphson* algorithm. If the optimum is interior and *f* is differentiable, it satisfies the necessary first order conditions that the gradient is zero

$$Df(\mathbf{x}^*) = \mathbf{0}. \tag{1.20}$$

Now apply a first order linear approximation to the gradient from some initial point **x**

$$\mathbf{0} = Df(\mathbf{x}^*) \approx Df(\mathbf{x}) + D^2 f(\mathbf{x})\left(\mathbf{x}^* - \mathbf{x}\right). \tag{21}$$

By rearranging terms we get an approximation to **x***

$$\mathbf{x}^* \approx \mathbf{x} - D^2 f(\mathbf{x})^{-1} Df(\mathbf{x}). \tag{22}$$

From this we can construct a search algorithm that hopefully makes better and better approximations

$$\mathbf{x}_{s+1} = \mathbf{x}_s - D^2 f(\mathbf{x}_s)^{-1} Df(\mathbf{x}_s). \tag{23}$$

If we don't have analytic solutions to the gradient and Hessian we can use numerical approximations, for example

$$\begin{aligned} \frac{\partial f(\mathbf{x})}{\partial x_1} &\approx \frac{f\left(\mathbf{x} + \begin{bmatrix} \varepsilon \\ \mathbf{0} \end{bmatrix}\right) - f(\mathbf{x})}{\varepsilon} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} &\approx \frac{\dfrac{\partial f\left(\mathbf{x} + \begin{bmatrix} 0 \\ \varepsilon \\ \mathbf{0} \end{bmatrix}\right)}{\partial x_1} - \dfrac{\partial f(\mathbf{x})}{\partial x_1}}{\varepsilon} \end{aligned} \tag{24}$$

which is called the forward difference method.

One should be very careful with these methods. The first problem is only local optima will be found with this method. Secondly, the method will surely converge only in a neighbourhood of **x*.** So often it is crucial for convergence which starting point is chosen.

$$\sum_{t=1}^{T} \ln f\left(y_t \middle| y_{t-1}\right). \tag{25}$$

Smoothed Probabilities

Above we have calculated the probabilities $P\left(s_t \middle| y_t\right)$. We may often be interested in also $P\left(s_t \middle| y_T\right)$. Since these probabilities use more information they will generally give better inference about the states. The smoothed state probabilities are also necessary for the EM algorithm.

We evaluate this "smoothed" probabilities in the following steps.

First we need the following lemma

$$P\left(s_t = j \middle| s_{t+1} = i, y_T\right) = P\left(s_t = j \middle| s_{t+1} = i, y_t\right) \tag{26}$$

The past state depends directly only on the current state. The "opposite" $P\left(s_{t+1} = j \middle| s_t = i, y_T\right) = P\left(s_{t+1} = j \middle| s_t = i, y_t\right)$ was an explicit assumption of Markov chains and the lemma seems quite intuitive. The proof is in Hamilton.

We also need another "lemma"

$$
\begin{aligned}
&P(s_t = j | s_{t+1} = i, y_t) \\
&= \frac{P(s_t = j, s_{t+1} = i | y_t)}{P(s_{t+1} = i | y_t)} \\
&= \frac{P(s_t = j | y_t) P(s_{t+1} = i | s_t = j)}{P(s_{t+1} = i | y_t)} \\
&= \frac{P(s_t = j |, y_t) p_{ji}}{P(s_{t+1} = i | y_t)}
\end{aligned} \tag{27}
$$

Now we start from the next to last period.

$$
\begin{aligned}
&P(s_{T-1} = j, s_T = i | y_T) \\
&= P(s_{T-1} = j | s_T = i, y_T) P(s_T = i | y_T).
\end{aligned} \tag{28}
$$

Now use the first lemma

$$
\begin{aligned}
&P(s_{T-1} = j, s_T = i | y_T) \\
&= P(s_{T-1} = j | s_T = i, y_{T-1}) P(s_T = i | y_T).
\end{aligned} \tag{29}
$$

Then we use the second lemma

$$
\begin{aligned}
&P(s_{T-1} = j, s_T = i | y_T) \\
&= \frac{P(s_{T-1} = j | y_{T-1}) p_{ji}}{P(s_T = i | y_{T-1})} P(s_T = i | y_T).
\end{aligned} \tag{30}
$$

The RHS of (30) we calculated when going through the 5 steps above. Then to get the smoothed probability for *T*-1 we just sum equation (30) over the states in *T*.

$$
P(s_{T-1} = j | y_T) = \sum_i P(s_{T-1} = j, s_T = i | y_T). \tag{31}
$$

Doing this for all states (31) gives a vector of smoothed state probabilities for *T*-1.

We can now go through the same steps for *T*-2 and on.

$$\begin{aligned} & P\left(s_{T-2}=j, s_{T-1}=i \middle| y_T\right) \\ & = P\left(s_{T-2}=j \middle| s_{T-1}=i, y_{T-2}\right) P\left(s_{T-1}=i \middle| y_T\right). \\ & = \frac{P\left(s_{T-2}=j \middle| y_{T-2}\right) p_{ji}}{P\left(s_{T-1}=i \middle| y_{T-2}\right)} P\left(s_{T-1}=i \middle|, y_T\right) \end{aligned} \tag{32}$$

where $P\left(s_{T-1}=i \middle| y_T\right)$ came from the previous iteration.

Diagnostics

The test we are going to perform are based on score tests. First define the score as

$$\mathbf{h}(t) \equiv \frac{\partial \ln f\left(y_t \middle| \mathbf{y}_{t-1}\right)}{\partial \theta} \tag{33}$$

evaluated at the true parameters θ and noting that we condition on the whole observed sequence of $\mathbf{y}_{t\text{-}1}, \ldots \mathbf{y}_0$ Now it is easy to understand that the score should be a martingale difference, i.e., it should have a zero expected value given all information available at *t*-1. Loosely speaking, we would otherwise expect to revise out estimate at *t* based on info available at *t*-1. We can show this formally by noting that

$$
\begin{aligned}
&\int_{-\infty}^{\infty} f\left(y_t \middle| \mathbf{y}_{t-1}\right) dy_t = 1 \\
&\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f\left(y_t \middle| \mathbf{y}_{t-1}\right) dy_t = 0 \\
&= \int_{-\infty}^{\infty} \frac{\partial f\left(y_t \middle| \mathbf{y}_{t-1}\right)}{\partial \theta} dy_t \\
&= \int_{-\infty}^{\infty} \frac{\partial \ln f\left(y_t \middle| \mathbf{y}_{t-1}\right)}{\partial \theta} f\left(y_t \middle| \mathbf{y}_{t-1}\right) dy_t \\
&= \int_{-\infty}^{\infty} \mathbf{h}(t) f\left(y_t \middle| \mathbf{y}_{t-1}\right) dy_t \\
&= E_{t-1}\mathbf{h}(t)
\end{aligned}
\tag{34}
$$

These scores are easily evaluated numerically and also often analytically along with the estimation of the Markov model. Based on (34) we can perform various tests. For example, we could run

$$
h_i(t) = \alpha_0 + \sum_j \alpha_j h_i(t-1) + \varepsilon_t \tag{35}
$$

for some combinations of *i* and *j*. For example, we could use the transition probability scores as both dependent and independent variables. Failure to pass this tests generates suspicion that the first order Markov assumption is invalid. We could also use transition probability scores as dependent and mean or standard deviation scores as independent. Failure here would indicate that the level or squared level of last periods realisation has a direct impact on the transition probability.

Using mean or standard deviation scores as dependent variables may give a way to find remaining AR or ARCH effects.

A problem with these tests is that we cannot expect the ε to be homoschedastic. We thus need to rely on Monte Carlo simulations to get a feel for the proper rejection levels. (See Hamilton, 1996).

The EM algorithm

Here follows an intuitive motivation for and an example of the EM algorithm in the regime switching model.

Assume we have two states $s_t \in \{0,1\}$. And that the innovations in the stochastic variable y_t are normally distributed. In state 0 with mean μ_0 and variance σ_0^2 and in state 1 with mean μ_1 and variance σ_1^2. The likelihood of an observation in state s_t is thus

$$f(y_t|s_t,\theta) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(y_t-\mu_s)^2}{2\sigma_s^2}} \tag{36}$$

Now assume the state was observable and deterministic. We could then write the lolikelihood function as

$$\begin{aligned} &L(\{y_t\};\theta) = \\ &\sum_{t=1}^{T}\left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_0^2 - \frac{(y_t-\mu_0)^2}{2\sigma_0^2}\right)D(s_t=0) \\ &+\sum_{t=1}^{T}\left(-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_1^2 - \frac{(y_t-\mu_1)^2}{2\sigma_1^2}\right)D(s_t=1) \end{aligned} \tag{37}$$

With first order conditions for state 0 mean and variance

$$\frac{\partial L(\{y_t\};\theta)}{\partial \mu_0} = \sum_{t=1}^{T}\left(\frac{(y_t-\mu_0)}{\sigma_0^2}\right)D(s_t=0)=0$$

$$\Rightarrow \frac{\sum_{t=1}^{T} y_t D(s_t=0)}{\sum_{t=1}^{T} D(s_t=0)} = \hat{\mu}_0 \tag{38}$$

and

$$\frac{\partial L(\{y_t\};\theta)}{\partial \sigma_0^2} = \sum_{t=1}^{T} -\left(\frac{1}{2\sigma_0^2}+\frac{(y_t-\hat{\mu}_0)^2}{2\sigma_0^4}\right)D(s_t=0)=0$$

$$\Rightarrow \sum_{t=1}^{T}\frac{(y_t-\hat{\mu}_0)^2}{\sigma_0^4}D(s_t=0) = \sum_{t=1}^{T}\frac{1}{2\sigma_0^2}D(s_t=0) \tag{39}$$

$$\frac{\sum_{t=1}^{T}(y_t-\hat{\mu}_0)^2 D(s_t=0)}{\sum_{t=1}^{T} D(s_t=0)} = \hat{\sigma}_0^2$$

and similarly for state 1.

Now the EM algorithm replaces the dummies in (38) and (39) with their smoothed probabilities.

$$\frac{\sum_{t=1}^{T} y_t P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)}{\sum_{t=1}^{T} P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)} = \hat{\mu}_0'$$

$$\frac{\sum_{t=1}^{T} (y_t - \mu_0)^2 P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)}{\sum_{t=1}^{T} P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)} = \sigma_0'^2$$

$$\frac{\sum_{t=2}^{T} P\left(s_t = 0, s_{t-1} = 0 \middle| y_T, \hat{\theta}\right) P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)}{\sum_{t=1}^{T} P\left(s_t = 0 \middle| y_T, \hat{\theta}\right)} = \hat{p}_{00}' \qquad (40)$$

and similarly for state 1 parameters. So for an initial guess we get an updated set of parameter estimates. Then we iterate on this until convergence.

It can be shown that such iterations always weakly increase the loglikelihood and that they converge to a maximum.

As you can see these approach reduces computation needs since in each iteration we do not need to compute Hessians if we use Newton Raphson. On the other hand we must compute smoothed probabilities. It turns out that usually the EM algorithm is much faster, in particular when analytical solutions to the parameter estimates are available and the parameter vector is large.

Extensions

The recursive estimation of the model is easily extended to allow for, for example:

Exogenous variables, e.g., dummies, a time trend.

State probabilities that are affected by levels or squared levels of exogenous or endogenous variables.

Vector processes.

More state variables