Regime Switching Models: An Example for a Stock Market Index

Erik Kole*

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam

April 2010

In this document, I discuss in detail how to estimate regime switching models with an example based on a US stock market index.

1 Specification

We assume that the returns on the US stock market index, Y_t, follow a distribution that depends on a latent process S_t. At each point in time, the process S_t is in one out of two regimes, which we indicate by $S_t = 0$ and $S_t = 1$. The return Y_t behaves according to

$$Y_t \sim \begin{cases} N(\mu_0, \sigma_0^2) & \text{if } S_t = 0 \\ N(\mu_1, \sigma_1^2) & \text{if } S_t = 1. \end{cases} \tag{1}$$

In both regimes, the return follows a normal distribution, though with different means and variances. We use the function f to denote the normal PDF,

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \tag{2}$$

Of course it is possible to have different distributions in regime 0 and 1.

The latent process S_t follows a first order Markov chain. This means that the probability for regime 0 to occur at time t depends solely on the regime at time $t-1$. We denote these transition probabilities by

$$p_{ij} = \Pr[S_t = i | S_{t-1} = j] \tag{3}$$

The transition probabilities for the departure states j should add up to one, i.e., $p_{00} + p_{10} = 1$ and $p_{01} + p_{11} = 1$. So, for a binary process S_t, we have two free parameters, p_{00} and p_{11}.

*Corresponding author. Address: Burg. Oudlaan 50, Room H11-13, P.O. Box 1738, 3000DR Rotterdam, The Netherlands, Tel. +31 10 408 12 58. E-mail addresses kole@ese.eur.nl.

We gather the transition probabilities in a transition matrix

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} p_{00} & 1 - p_{11} \\ 1 - p_{00} & p_{11} \end{pmatrix}. \tag{4}$$

Since the whole process S_t is unobserved, this also applies to the initial regime S_1. We introduce a separate parameter ζ for the probability that the first regime occurs,

$$\zeta = \Pr[S_1 = 0]. \tag{5}$$

Naturally, we have $\Pr[S_1 = 1] = 1 - \zeta$. Because no conditional information on S_0 is available, we cannot directly use the transition matrix to determine this probability, and we need the extra parameter. This last parameter can be estimated, but also specified exogenously. We assume in this document that the parameter is estimated.

2 Inference on S_t

The process S_t is latent, which means that we will never know for sure which regime prevailed at a certain point in time. However, we can use the information from the current and past observations, combined with the distributions and transition probabilities to make an *inference* on $\Pr[S_t = 0|y_t, y_{t-1}, \ldots, y_1]$. We accomplish this by using Bayes' rule,

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]}.$$

For the inference of the regime at time $t = 1$, this means

$$\begin{aligned}
\Pr[S_1 = 0|Y_1 = y_1] &= \frac{\Pr[Y_1 = y_1|S_1 = 0] \cdot \Pr[S_1 = 0]}{\Pr[Y_1 = y_1]} \\
&= \frac{\Pr[Y_1 = y_1|S_1 = 0] \cdot \Pr[S_1 = 0]}{\Pr[Y_1 = y_1|S_1 = 0] \cdot \Pr[S_1 = 0] + \Pr[Y_1 = y_1|S_1 = 1] \cdot \Pr[S_1 = 1]} \\
&= \frac{f(y_1; \mu_0, \sigma_0^2) \cdot \zeta}{f(y_1; \mu_0, \sigma_0^2) \cdot \zeta + f(y_1; \mu_1, \sigma_1^2) \cdot (1 - \zeta)}.
\end{aligned}$$

In the second equality, we use conditioning again, because conditional on the regime the distribution of Y_1 is given. We make the distributions explicit in the third equality. In a similar way, we find an expression for $\Pr[S_1 = 1|Y_1 = y_1]$, but we can also compute this using $\Pr[S_1 = 1|Y_1 = y_1] = 1 - \Pr[S_1 = 0|Y_1 = y_1]$.

After computing the inferences for the regimes at time 1, we can use them to make a *forecast* for the regime distribution at time 2,

$$\begin{aligned}
\Pr[S_2 = 0|Y_1 = y_1] = {} & \Pr[S_2 = 0|S_1 = 0, Y_1 = y_1] \cdot \Pr[S_1 = 0|Y_1 = y_1] + \\
& \Pr[S_2 = 0|S_1 = 1, Y_1 = y_1] \cdot \Pr[S_1 = 1|Y_1 = y_1] \\
& = \Pr[S_2 = 0|S_1 = 0] \cdot \Pr[S_1 = 0|Y_1 = y_1] + \\
& \Pr[S_2 = 0|S_1 = 1] \cdot \Pr[S_1 = 1|Y_1 = y_1] \\
& = p_{00}\Pr[S_1 = 0|Y_1 = y_1] + p_{01}\Pr[S_1 = 1|Y_1 = y_1].
\end{aligned}$$

In the first equality we condition on the regime at time 1. In the second equality we use the fact that S_t follows a first order Markov chain independent of the process Y_t. Again, we can similarly derive $\Pr[S_2 = 1|Y_1 = y_1]$ or use $\Pr[S_2 = 1|Y_1 = y_1] = 1 - \Pr[S_2 = 0|Y_1 = y_1]$.

The steps of calculating inference and forecast probabilities define a recursion. Based on the forecast probabilities for time 2 and the observation y_2 we can calculate inference probabilities for the regime at time 2. In turn, we use these inferences for forecasts for the regime at time 3. We can write these recursions compacter by using vector-matrix notation. We use $\boldsymbol{\xi}_{t|t} = \Pr[S_t|y_t, y_{t-1}, \ldots, y_1]$ to denote the vector of inferences probabilities at time t, and $\boldsymbol{\xi}_{t+1|t} = \Pr[S_{t+1}|y_t, y_{t-1}, \ldots, y_1]$ for the forecast probabilities at time t, using information up to time t. We gather the densities of observation y_t conditional on the regimes in a vector $\boldsymbol{f}_t$. We can construct the series of inference and forecast probabilities by the recursion

$$\boldsymbol{\xi}_{t|t} = \frac{1}{\boldsymbol{\xi}'_{t|t-1}\boldsymbol{f}_t}\boldsymbol{\xi}_{t|t-1} \odot \boldsymbol{f}_t \tag{6}$$

$$\boldsymbol{\xi}_{t+1|t} = \boldsymbol{P}\boldsymbol{\xi}_{t|t}, \tag{7}$$

where $\odot$ indicates element-by-element multiplication. We call this the filter recursion.

It is also possible to determine the probability of the occurrence of a specific regime at time t, using all available information, i.e., information before and after time t, which we call smoothed inference probabilities. These probabilities denoted by $\boldsymbol{\xi}_{t|T}$ can also be calculated by recursion,

$$\boldsymbol{\xi}_{t|T} = \boldsymbol{\xi}_{t|t} \odot \left(\boldsymbol{P}'(\boldsymbol{\xi}_{t+1|T} \div \boldsymbol{\xi}_{t+1|t})\right), \tag{8}$$

where we use the inference and forecast probabilities (see Kim, 1994, §2.2, for a derivation). We use the smoothed inference probabilities mostly to show how the regimes are identified. This recursion is called the smoother recursion.

3 Estimation

We can estimate the parameters of the regime switching models using a maximum likelihood approach. As with other conditional models such as ARMA- or GARCH-models, the likelihood function will take a conditional form, too. We gather the parameters of the model in a vector $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2, p_{00}, p_{11}, \zeta)'$. The conditional likelihood function is given by

$$\mathcal{L}(y_1, y_2, \ldots, y_T; \boldsymbol{\theta}) = \prod_{t=1}^{T} \Pr[Y_t = y_t|y_{t-1}, y_{t-2}, \ldots y_1]. \tag{9}$$

Conditioning on the regime at time t, we find

$$\begin{aligned}\Pr[Y_t = y_t|y_1, y_2, \dots y_{t-1}] =& \Pr[Y_t = y_t|S_t = 0, y_1, y_2, \dots y_{t-1}] \cdot \Pr[S_t = 0|y_1, y_2, \dots y_{t-1}]+\\ &\Pr[Y_t = y_t|S_t = 1, y_1, y_2, \dots y_{t-1}] \cdot \Pr[S_t = 1|y_1, y_2, \dots y_{t-1}]\\ =& \Pr[Y_t = y_t|S_t = 0] \cdot \xi_{t|t-1,0} + \Pr[Y_t = y_t|S_t = 1] \cdot \xi_{t|t-1,1}\\ =& \boldsymbol{\xi}'_{t|t-1}\boldsymbol{f}_t\end{aligned}$$

In the second equality, we use the information that the distribution of $Y_t|S_t$ does not depend on further prior realizations. The conditional log likelihood function can thus be calculated as

$$\ell(y_1, y_2, \dots, y_T; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log(\boldsymbol{\xi}'_{t|t-1}\boldsymbol{f}_t), \tag{10}$$

which follows as a byproduct of the filter recursion.

Straightforward maximum likelihood estimation implies maximizing (10) as a function of $\boldsymbol{\theta}$. Because of the filter recursion, the log likelihood function exhibits a complicated structure with many local optima. Optimizing this function may therefore be computationally demanding. Therefore, we will use a special optimization algorithm, called the Expectation-Maximization (EM) algorithm of Dempster et al. (1977).

3.1 The Expectation-Maximization Algorithm

Suppose that we could actually observe the realizations of the latent process S_t, and we would have a set $\{s_1, s_2, \dots, s_T\}$ similar to the set $\{y_1, y_2, \dots, y_T\}$. To simplify notation, we write $\mathcal{S}_t = \{s_1, s_2, \dots, s_t\}$ and $\mathcal{Y}_t = \{y_1, y_2, \dots, y_t\}$. The realization of S_t is either zero or one, so it corresponds with a draw from a Bernoulli distribution. We find the density of the combination (y_t, s_t) conditional on past observations as

$$\begin{aligned}\Pr[Y_t = y_t, S_t = s_t|\mathcal{Y}_{t-1}, \mathcal{S}_{t-1}; \boldsymbol{\theta}] =& \Pr[Y_t = y_t|\mathcal{S}_t; \boldsymbol{\theta}] \Pr[S_t = s_t|\mathcal{S}_{t-1}; \boldsymbol{\theta}]\\ =& \begin{cases} f(y_t; \mu_0, \sigma_0^2)p_{00} & \text{if } s_t = 0, s_{t-1} = 0\\ f(y_t; \mu_0, \sigma_0^2)(1-p_{11}) & \text{if } s_t = 0, s_{t-1} = 1\\ f(y_t; \mu_1, \sigma_1^2)(1-p_{00}) & \text{if } s_t = 1, s_{t-1} = 0\\ f(y_t; \mu_1, \sigma_1^2)p_{11} & \text{if } s_t = 1, s_{t-1} = 1\end{cases}\\ =& \big(f(y_t; \mu_0, \sigma_0^2)p_{00}\big)^{(1-s_t)(1-s_{t-1})} \times\\ &\big(f(y_t; \mu_0, \sigma_0^2)(1-p_{11})\big)^{(1-s_t)s_{t-1}} \times\\ &\big(f(y_t; \mu_1, \sigma_1^2)(1-p_{00})\big)^{s_t(1-s_{t-1})} \times\\ &\big(f(y_t; \mu_1, \sigma_1^2)p_{11}\big)^{s_t s_{t-1}}.\end{aligned} \tag{11}$$

We see that the density of (y_t, s_t) combines the fact that conditionally, y_t follows a normal distribution, with the fact that s_t follows a Bernoulli distribution, conditionally on its previous realization s_{t-1}.

When we construct the log likelihood function of the joint observations $(\mathcal{Y}_T, \mathcal{S}_T)$, we need the log of (11)

$$
\begin{aligned}
&\log \Pr[Y_t = y_t, S_t = s_t | \mathcal{Y}_{t-1}, \mathcal{S}_{t-1}; \boldsymbol{\theta}] \\
=&\log\left(f\left(y_t; \mu_0, \sigma_0^2\right) p_{00}\right) \cdot (1 - s_t) \cdot (1 - s_{t-1}) + \\
&\log\left(f\left(y_t; \mu_0, \sigma_0^2\right) (1 - p_{11})\right) \cdot (1 - s_t) \cdots_{t-1} + \\
&\log\left(f\left(y_t; \mu_1, \sigma_1^2\right) (1 - p_{00})\right) \cdot s_t \cdot (1 - s_{t-1}) + \\
&\log\left(f\left(y_t; \mu_1, \sigma_1^2\right) p_{11}\right) \cdot s_t \cdot s_{t-1} \\
=&(1 - s_t) \log f\left(y_t; \mu_0, \sigma_0^2\right) + s_t \log f\left(y_t; \mu_1, \sigma_1^2\right) + \\
&(1 - s_t)(1 - s_{t-1}) \log p_{00} + (1 - s_t)s_{t-1} \log(1 - p_{11}) + \\
&s_t(1 - s_{t-1}) \log(1 - p_{00}) + s_t s_{t-1} \log p_{11}
\end{aligned}
$$

A small alteration must be made for the density of $\Pr[Y_1 = y_1, S_1 = s_1; \boldsymbol{\theta}]$, since no history will be available there. So, instead of the Markov chain parameters p_{00} and p_{11} we find an expression with the parameter ζ,

$$
\Pr[Y_1 = y_1, S_1 = s_1; \boldsymbol{\theta}] = \left(f(y_t; \mu_0, \sigma_0^2)\zeta\right)^{(1-s_1)} \left(f(y_t; \mu_1, \sigma_1^2)\zeta\right)^{s_1}.
$$

Now, we can simply construct the log likelihood for $(\mathcal{Y}_T, \mathcal{S}_T)$ as

$$
\begin{aligned}
\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) = &\sum_{t=1}^{T} \Big((1 - s_t) \log f(y_t; \mu_0, \sigma_0^2) + s_t \log f(y_t; \mu_1, \sigma_1^2)\Big) + \\
&\sum_{t=2}^{T} \Big((1 - s_t)(1 - s_{t-1}) \log p_{00} + (1 - s_t)s_{t-1} \log(1 - p_{11}) + \\
&\quad s_t(1 - s_{t-1}) \log(1 - p_{00}) + s_t s_{t-1} \log p_{11}\Big) + \\
&\quad (1 - s_1) \log \zeta + s_1 \log(1 - \zeta)
\end{aligned} \tag{12}
$$

This log likelihood function would be much easier to optimize than the actual log likelihood function in (10), because (12) does not exhibit a recursive relation. However, we cannot observe $\mathcal{S}_t$.

The EM-algorithm proposes to base the estimation on (12). Because we do not have actual observations on S_t, the EM-algorithm maximizes the expectation of the log likelihood function in (12) based on the complete data that we do observe, $\mathcal{Y}_T$. So, instead of working with s_t, we work with the expectation of S_t conditional on the data and the parameters,

$$
\mathrm{E}[S_t | \mathcal{Y}_T; \boldsymbol{\theta}] = \Pr[S_t = 0 | \mathcal{Y}_T; \boldsymbol{\theta}] \cdot 0 + \Pr[S_t = 1 | \mathcal{Y}_T; \boldsymbol{\theta}] \cdot 1 = \Pr[S_t = 1 | \mathcal{Y}_T; \boldsymbol{\theta}]. \tag{13}
$$

The last probability is a smoothed inference probability as in (8). Similarly, we find

$$
\mathrm{E}[S_t S_{t-1} | \mathcal{Y}_T; \boldsymbol{\theta}] = \Pr[S_t = S_{t-1} = 1 | \mathcal{Y}_T; \boldsymbol{\theta}]. \tag{14}
$$

This approach would almost retain the attractive structure of the log likelihood function in (12). Almost, as the expectations of S_t and S_tS_{t-1} depend on $\boldsymbol{\theta}$ and are calculated again via the recursion in (8). The trick of the EM-algorithm is to treat the expectation part and the maximization separately. So, for a given parameter vector $\boldsymbol{\theta}$, the expectations in (13) and (14) are calculated. Then, these expectations are treated as given, and a new parameter vector $\boldsymbol{\theta}^*$ is calculated which maximizes the expected log likelihood function. Of course, this new parameter vector gives rise to other expectations, which in turn lead to a new parameter vector. So, instead of one direct maximum likelihood estimation, we conduct a series of expectation maximization steps, which produce a series of parameter estimates $\boldsymbol{\theta}^{(k)}$

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathrm{E}\left[\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) | \mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}\right]. \tag{15}$$

Dempster et al. (1977) and Hamilton (1990) show that this sequence of $\boldsymbol{\theta}^{(k)}$ converges and produces a maximum of (10). As always, this maximum can be local, and may depend on starting values $\boldsymbol{\theta}^{(0)}$.

3.2 The Maximization Step

We now look at the maximization step in more detail. Our starting point is the likelihood function in (12), for which we calculate the expectation conditional on the data and parameters $\boldsymbol{\theta}^{(k-1)}$,

$$\ell_{\mathrm{EM}}\big(\mathcal{Y}_T; \boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}\big) = \mathrm{E}\left[\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) | \mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}\right] \tag{16}$$

The updated parameters $\boldsymbol{\theta}^{(k)}$ maximize this expected log likelihood function, so they satisfy the first order conditions

$$\left.\frac{\partial \ell_{\mathrm{EM}}\big(\mathcal{Y}_T; \boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}\big)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = 0. \tag{17}$$

Taking a closer look at (12), we see that the log likelihood function can be split in terms that exclusively relate to specific parameters. The parameters of the distribution for the first regime μ_0 and σ_0^2 are only related to the first term, and the parameters of the distribution for the second regime only to the second. The transition probability p_{00} is related to the third and fifth term, and so on. So differentiation will produce relatively simply conditions.

We first look at differentiating (16) with respect to μ_0. We will use $\xi_{t|T,0}^{(k-1)}$ to denote $\Pr[S_t = 0|\mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}] = 1 - \mathrm{E}[S_t|\mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}]$, which is the smoothed inference probability that we find when we apply the filter and smoother recursions in (6)–(8) with parameters

$\boldsymbol{\theta}^{(k-1)}$. We find

$$\begin{aligned}
\frac{\partial \ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial \mu_0} &= \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \log f(y_t;\mu_0,\sigma_0^2)}{\partial \mu_0} \\
&= \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(-\frac{1}{2}\log 2\pi - \log\sigma_0 - \frac{1}{2}\frac{(y_t-\mu_0)^2}{\sigma_0^2}\right)}{\partial \mu_0} \\
&= \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \frac{(y_t-\mu_0)}{\sigma_0^2}
\end{aligned} \tag{18}$$

For the optimal $\mu_0^{(k)}$ this expression equals zero, which means that we find

$$\mu_0^{(k)} = \frac{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} y_t}{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)}}. \tag{19}$$

This estimate for μ_0 can be interpreted as a weighted average of the observations, where the smoothed inference probabilities for regime 0 serve as weights. It is a clear extension of the normal maximum likelihood estimator for the mean of a normal distribution. For $\mu_1^{(k)}$ we find a similar expression, with $\xi_{t|T,1}^{(k-1)}$ instead of $\xi_{t|T,0}^{(k-1)}$.

Next we consider the estimates for σ_0^2. Differentiation yields

$$\begin{aligned}
\frac{\partial \ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial \sigma_0} &= \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \log f(y_t;\mu_0,\sigma_0^2)}{\partial \sigma_0} \\
&= \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(-\frac{1}{2}\log 2\pi - \log\sigma_0 - \frac{1}{2}\frac{(y_t-\mu_0)^2}{\sigma_0^2}\right)}{\partial \sigma_0} \\
&= \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(\frac{(y_t-\mu_0)^2}{\sigma_0^3} - \frac{1}{\sigma_0}\right).
\end{aligned} \tag{20}$$

The optimal $\sigma_0^{(k)}$ sets this expression to zeros, so

$$\sigma_0^{(k)} = \sqrt{\frac{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(y_t - \mu_0^{(k)}\right)^2}{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)}}}, \tag{21}$$

which is again a weighted average.

In a similar way we can derive the estimates for p_{00} and p_{11}. Before we derive these estimates, note that

$$\begin{aligned}
&\mathrm{E}\big[(1-S_t)(1-S_{t-1})|\mathcal{Y}_T;\boldsymbol{\theta}\big]\\
=&1-\mathrm{E}\big[S_t|\mathcal{Y}_T;\boldsymbol{\theta}\big]-\mathrm{E}\big[S_{t-1}|\mathcal{Y}_T;\boldsymbol{\theta}\big]+\mathrm{E}\big[S_tS_{t-1}|\mathcal{Y}_T;\boldsymbol{\theta}\big]\\
=&1-\Pr\big[S_t=1|\mathcal{Y}_T;\boldsymbol{\theta}\big]-\Pr\big[S_t=1|\mathcal{Y}_T;\boldsymbol{\theta}\big]+\Pr\big[S_t=S_{t-1}=1|\mathcal{Y}_T;\boldsymbol{\theta}\big]\\
=&\Pr\big[S_t=S_{t-1}=0|\mathcal{Y}_T;\boldsymbol{\theta}\big]
\end{aligned}$$

and similarly $\mathrm{E}\big[S_t(1-S_{t-1})|\mathcal{Y}_T;\boldsymbol{\theta}\big]=\Pr\big[S_t=1,S_{t-1}=0|\mathcal{Y}_T;\boldsymbol{\theta}\big]$ and $\mathrm{E}\big[(1-S_t)S_{t-1}|\mathcal{Y}_T;\boldsymbol{\theta}\big]=\Pr\big[S_t=0,S_{t-1}=1|\mathcal{Y}_T;\boldsymbol{\theta}\big]$. These probabilities can be calculated with a slight modification of the recursion in (8),

$$\Pr[S_{t+1}=i,S_t=j|\mathcal{Y}_T;\boldsymbol{\theta}^{(k-1)}]=\tilde{p}_{ij,t+1}=\xi_{t|t,j}\cdot\frac{\xi_{t+1|T,i}}{\xi_{t+1|t,i}}p_{ij}^{(k-1)} \tag{22}$$

The derivative for p_{00} is given by

$$\begin{aligned}
\frac{\partial\ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial p_{00}}&=\frac{\partial\sum_{t=2}^{T}\tilde{p}_{00,t}\log p_{00}+\tilde{p}_{10,t}\log(1-p_{00})}{\partial p_{00}}\\
&=\sum_{t=2}^{T}\left(\frac{\tilde{p}_{00,t}}{p_{00}}-\frac{\tilde{p}_{10,t}}{1-p_{00}}\right).
\end{aligned} \tag{23}$$

Setting this expression to zero implies

$$p_{00}^{(k)}=\frac{\sum_{t=2}^{T}\tilde{p}_{00,t}}{\sum_{t=2}^{T}\left(\tilde{p}_{00,t}+\tilde{p}_{10,t}\right)}=\frac{\sum_{t=2}^{T}\tilde{p}_{00,t}}{\sum_{t=2}^{T}\xi_{t-1|T,0}}. \tag{24}$$

This can be generalized to

$$p_{ij}^{(k)}=\frac{\sum_{t=2}^{T}\tilde{p}_{ij,t}}{\sum_{t=2}^{T}\xi_{t-1|T,j}}, \tag{25}$$

which corresponds with (3.45) in Franses and van Dijk (2000).

Finally, we consider the estimate for the ζ parameter, which is easy to derive. The derivative of interest is

$$\begin{aligned}
\frac{\partial\ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial\zeta}&=\frac{\partial\left(\xi_{1|T,0}\log\zeta+\xi_{1|T,1}\log(1-\zeta)\right)}{\partial\zeta}\\
&=\frac{\xi_{1|T,0}}{\zeta}-\frac{\xi_{1|T,1}}{1-\zeta}.
\end{aligned} \tag{26}$$

Setting this expression to zero we find

$$\zeta^{(k)}=\xi_{t|T,0}^{(k-1)}. \tag{27}$$

3.3 Remarks

1. The EM-algorithm needs starting values $\boldsymbol{\theta}^{(0)}$. In principle, these starting values can be picked at random, as long as they are feasible, i.e., positive volatility and probabilities between zero and one. It is advisable to make sure that the distribution parameters for regime 0 differ substantially from those for regime 1. For example, take the volatility for regime 1 three or four times that for regime 0. Regimes tend to be persistent, so set the transition probabilities at a high value of 0.9, say

2. The EM-algorithm converges and maximizes the likelihood. This means that each maximization step in the EM-algorithm should yield an improvement. In other words, for each new set of parameters $\boldsymbol{\theta}^{(k)}$, the log likelihood function in (10) should increase. In implementing the algorithm, an important control mechanism is whether $\ell\big(\mathcal{Y}_T;\boldsymbol{\theta}(k)\big) > \ell\big(\mathcal{Y}_T;\boldsymbol{\theta}^{(k-1)}\big)$. If not, the EM-algorithm is not implemented correctly.

3. Each step in the EM-algorithm yields an improvement in the likelihood function. This improvement will get smaller and smaller, with parameters that also do not change very much. So, you have to specify a stopping criterion, which is best formulated for the increase in likelihood falling below a threshold.

4 An example

In the example we look at weekly excess returns on the MSCI US Stock Market Index. For each week, I have calculated the log return on the index, from which I have subtracted the 1-week risk free rate. The first return is for January 2, 1980 and the last for July 1, 2009. In total we have 1540 observations (see Kole and van Dijk, 2010, for more details on the data). The data is available in the file `RSExample_MSCIUS.xls`. The returns are given in %.

4.1 Inferences

First, we look at the inferences that we make for a given set of parameters. As values for the parameters we take

$$\begin{array}{llll} \mu_0 = 0.04 & \sigma_0 = 1 & p_{11} = 0.80 & \zeta = 0.50 \\ \mu_1 = -0.04 & \sigma_1 = 4 & p_{22} = 0.80 & \end{array}$$

The means and volatilities are based on the overall sample mean, which was close to zero, and the overall sample variance which was around two.

In Table 1 we see the first ten forecast, inference and smoothed inference probabilities. The first forecast probabilities are given by ζ and $1-\zeta$. Based on the first return of -1.01923, the inference probabilities are calculated. This return is relatively close to zero, and fits better with the first regime (low volatility) than the second regime (high volatility). Therefore the inference probability for state 0 is higher than for state 1. Because of the

Table 1: Inferences for the first ten returns.

observation	return	forecast probabilities $S_t = 0$	forecast probabilities $S_t = 1$	inference probabilities $S_t = 0$	inference probabilities $S_t = 1$	smoothed inf. probabilities $S_t = 0$	smoothed inf. probabilities $S_t = 1$
1	−1.01923	0.50000	0.50000	0.70167	0.29833	0.51467	0.48533
2	2.64830	0.62100	0.37900	0.21490	0.78510	0.27057	0.72943
3	1.54639	0.32894	0.67106	0.40549	0.59451	0.45034	0.54966
4	2.02344	0.44329	0.55671	0.33727	0.66273	0.51982	0.48018
5	0.96257	0.40236	0.59764	0.64486	0.35514	0.72967	0.27033
6	0.04977	0.58691	0.41309	0.85040	0.14960	0.73656	0.26344
7	1.81177	0.71024	0.28976	0.69432	0.30568	0.40332	0.59668
8	−2.47153	0.61659	0.38341	0.24830	0.75170	0.07637	0.92363
9	−4.24477	0.34898	0.65102	0.00038	0.99962	0.00018	0.99982
10	−1.69100	0.20023	0.79977	0.19599	0.80401	0.05800	0.94201

This tables shows the first ten returns with their forecast probabilities, inference probabilities and smoothed inference probabilities. The inferences are based on the two-state regime switching model specified in Sec. 1. The parameters values are $\mu_0 = 0.04$, $\sigma_0 = 1$, $\mu_1 = -0.04$ $\sigma_1 = 4$, $p_{11} = 0.80$, $p_{22} = 0.80$ and $\zeta = 0.50$.

persistence of the regimes (p_{11} and p_{22} are high), the forecast probability for state 0 at time 2, is higher than the 0.5 at time 1. Returns at time 2, 3 and 4 match better with the high volatility regime (inference probabilities for regime 2 exceed 0.5). Consequently, when we smooth the series of inference probabilities, the probability for regime 0 at time 1 goes down, from 0.70167 to 0.51467.

4.2 Estimation

We can use the parameters we picked in the previous subsection to start the EM-algorithm to estimate the model parameters. We set the stopping criterion at an increase in the log likelihood function in (10) below 10^{-8}. In Table 2 we show how the EM algorithm proceeds. We see that the likelihood increases with every iteration. The EM-algorithm needs 48 steps in 0.719 seconds to converge to the optimal solution in this case.

In Table 3 we report the forecast, inference and smoothed inference probabilities for the first ten returns, based on the parameters estimates produced by the EM-algorithm. Compared to Table 1, we see the regimes are better defined now: the probabilities are either close to zero or to one. The inference probabilities signal a possible switch for the return after 9 weeks, where the probability for regime 2 increases above 0.5. It is still close to 0.5, so based on the 9 weeks of information the regime switching models does not produce certain inferences about the switch. Using all information, the inference is more certain for regime 2, and dates the switch already in week 8.

In Figure 1, we see the smoothed inference probabilities for regime 0 over time. This low volatility regime prevails during prolonged periods of time, but we also see clear periods identified as exhibiting high volatility, notably around the crash of October 1987, the Asian crisis (1997), the Ruble crisis (1998), the burst of the IT-bubble after 2001 and the credit crisis in 2007-2008.

Table 2: Steps of the EM-algorithm

	starting values	iteration 1	2	3	optimal solution
μ_0	0.0400	0.1426	0.1980	0.2240	0.1573
σ_0	1.0000	1.1445	1.2182	1.2645	1.5594
μ_1	-0.0400	-0.1262	-0.1887	-0.2324	-0.2988
σ_1	4.0000	3.1417	3.0916	3.1030	3.4068
p_{11}	0.8000	0.8222	0.8345	0.8532	0.9770
p_{22}	0.8000	0.7899	0.8072	0.8195	0.9484
ζ	0.5000	0.5147	0.5585	0.6501	1.0000
$\ell(\mathcal{Y}_T;\boldsymbol{\theta})$	-3423.5840	-3352.8306	-3343.2509	-3337.7226	-3310.2279

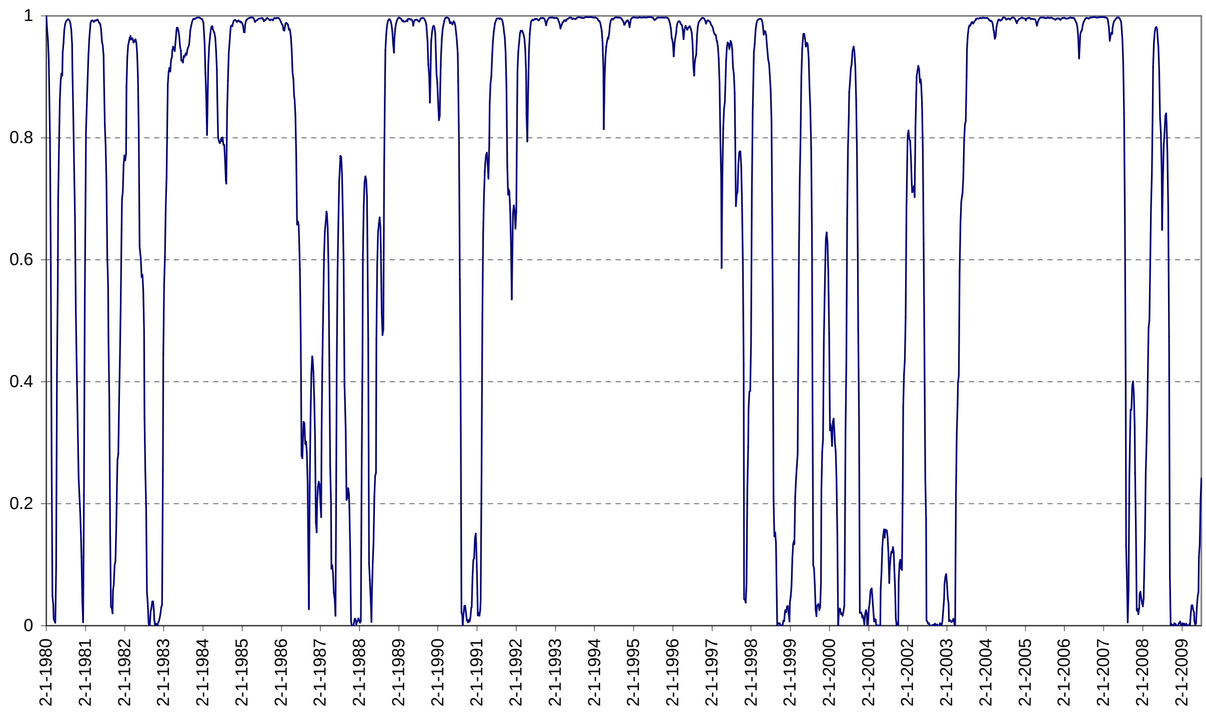
This table shows the steps of the EM-algorithm, applied to the full sample. Starting values for the parameters are $\mu_0 = 0.04$, $\sigma_0 = 1$, $\mu_1 = -0.04$ $\sigma_1 = 4$, $p_{11} = 0.80$, $p_{22} = 0.80$ and $\zeta = 0.50$. The algorithm stops when the improvement in the log likelihood function falls below 10^{-8}. We show the parameters after the first three iterations, and the optimal values. For each parameter set we calculate the value of the log likelihood function in (10).

Table 3: Inferences for the first ten returns, based on estimated parameters.

observation	return	forecast probabilities $S_t = 0$	$S_t = 1$	inference probabilities $S_t = 0$	$S_t = 1$	smoothed inf. probabilities $S_t = 0$	$S_t = 1$
1	-1.01923	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
2	2.64830	0.97697	0.02303	0.97411	0.02589	0.97756	0.02244
3	1.54639	0.95301	0.04699	0.97184	0.02816	0.95963	0.04037
4	2.02344	0.95091	0.04909	0.96308	0.03692	0.92842	0.07158
5	0.96257	0.94281	0.05719	0.97123	0.02877	0.88600	0.11400
6	0.04977	0.95035	0.04965	0.97671	0.02329	0.79482	0.20518
7	1.81177	0.95542	0.04458	0.96998	0.03002	0.58738	0.41262
8	-2.47153	0.94919	0.05081	0.92354	0.07646	0.26443	0.73557
9	-4.24477	0.90622	0.09378	0.43437	0.56563	0.04898	0.95103
10	-1.69100	0.45357	0.54643	0.49407	0.50593	0.03344	0.96657

This tables shows the first ten returns with their forecast probabilities, inference probabilities and smoothed inference probabilities. The inferences are based on the two-state regime switching model specified in Sec. 1. The parameters are estimated with the EM-algorithm and reported in Table 2.

Figure 1: Smoothed Inference Probability for Regime 0



This figure shows the smoothed inference probabilities for regime 0 over time for the US stock market. The probabilities are constructed using the filter recursion in (6) and (7) and the smoother recursion of Kim (1994) in (8). The parameters are estimated with the EM-algorithm and reported in Table 2.

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38.

Franses, P. H. and van Dijk, D. (2000). *Non-Linear Time Series Models in Empirical Finance*. Cambridge University Press, Cambridge, UK.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70.

Kim, C.-J. (1994). Dynamic linear models with markov-switching. *Journal of Econometrics*, 60(1):1–22.

Kole, E. and van Dijk, D. J. C. (2010). How to predict bull and bear markets? Working paper, Econometric Institute, Erasmus University Rotterdam, The Netherlands.