Markov Switching Models: An Example for a Stock Market Index

Erik Kole*

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam

This version: May 2019
First version: April 2010

Abstract

In this document, I discuss in detail how to estimate Markov regime switching models with an example based on a US stock market index. See for example Kole and Dijk (2017) for an application.

Key words: Markov switching, Expectation Maximization, bull and bear markets
JEL classification: C51, C58, A23

1 Specification

We assume that the asset return Y_t follows 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

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

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

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 ergodic 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}. We gather the transition probabilities in a transition matrix[1]

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

Since the whole process S_t is unobserved, so is 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 about S_1 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

[1] Formulating the transition matrix such that the columns sum to one is more convenient here.

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 events A and B with $\Pr[B] \neq 0$. 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,

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

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 the inference about and forecast for the states define a recursion. Based on the regime forecast for time 2 and the observation y_2 we can calculate the inference about the regime at time 2. In turn, we use these inferences for forecasts for the regime at time 3. We can write these recursions more compact by using vector-matrix notation. We use

$$\boldsymbol{\xi}_{t|t} \equiv \begin{pmatrix} \Pr[S_t = 0|y_t, y_{t-1}, \ldots, y_1] \\ \Pr[S_t = 1|y_t, y_{t-1}, \ldots, y_1] \end{pmatrix} \tag{6}$$

to denote the vector of inferences about the regimes at time t, and

$$\boldsymbol{\xi}_{t+1|t} \equiv \begin{pmatrix} \Pr[S_{t+1} = 0 | y_t, y_{t-1}, \ldots, y_1] \\ \Pr[S_{t+1} = 1 | y_t, y_{t-1}, \ldots, y_1] \end{pmatrix} \tag{7}$$

for the regime forecasts at time $t+1$, using information up to time t. We gather the densities of observation y_t conditional on the regimes in a vector

$$\boldsymbol{f}_t \equiv \begin{pmatrix} f(y_t; \mu_0, \sigma_0^2) \\ f(y_t; \mu_1, \sigma_1^2) \end{pmatrix}. \tag{8}$$

We can construct the series of inference and forecast probabilities by the recursion

$$\boldsymbol{\xi}_{t|t} = \frac{1}{\boldsymbol{\xi}'_{t|t-1}\boldsymbol{f}_t} \boldsymbol{\xi}_{t|t-1} \odot \boldsymbol{f}_t \tag{9}$$

$$\boldsymbol{\xi}_{t+1|t} = \boldsymbol{P}\boldsymbol{\xi}_{t|t}, \tag{10}$$

where $\odot$ indicates element-by-element multiplication. This recursion is called the Hamilton filter (see Hamilton, 1994). It starts with $\boldsymbol{\xi}_{1|0} = (\zeta, 1-\zeta)'$.

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 inferences. These inferences

$$\boldsymbol{\xi}_{t|T} \equiv \begin{pmatrix} \Pr[S_t = 0 | y_T, y_{T-1}, \ldots, y_1] \\ \Pr[S_t = 1 | y_T, y_{T-1}, \ldots, y_1] \end{pmatrix} \tag{11}$$

can be calculated by the recursion,

$$\boldsymbol{\xi}_{t|T} = \boldsymbol{\xi}_{t|t} \odot \big(\boldsymbol{P}'(\boldsymbol{\xi}_{t+1|T} \div \boldsymbol{\xi}_{t+1|t})\big), \tag{12}$$

where we use the inferences and forecasts, and $\div$ denotes element-wise division (see Kim, 1994, §2.2, for a derivation). This recursion is sometimes called the Kim-smoother. Whereas the Hamilton filter runs forward in time, the Kim filter runs backwards starting with $\boldsymbol{\xi}_{T|T}$.

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 vectors $\boldsymbol{\lambda} \equiv (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2)'$, $\boldsymbol{\rho} \equiv (p_{00}, p_{11}, \zeta)'$ and $\boldsymbol{\theta} \equiv (\boldsymbol{\lambda}', \boldsymbol{\rho}')'$. To simplify notation, we write $\mathcal{Y}_t \equiv \{y_t, y_{t-1}, \ldots, y_1\}$ for the set of observations up to t.

The likelihood function is given by

$$\mathcal{L}(\mathcal{Y}_T; \boldsymbol{\theta}) = \prod_{t=1}^{T} \Pr[Y_t = y_t | \mathcal{Y}_{t-1}], \tag{13}$$

with $\mathcal{Y}_0 = \emptyset$ so $\Pr[Y_t = y_1 | \mathcal{Y}_0] = \Pr[Y_t = y_1]$. Conditioning on the regime at time t, we find

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

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

$$\ell(y_1, y_2, \ldots, y_T; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log(\boldsymbol{\xi}_{t|t-1}' \boldsymbol{f}_t), \tag{14}$$

which follows as a byproduct of the filter recursion.

Straightforward maximum likelihood estimation implies maximizing (14) 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 $\mathcal{S}_t \equiv \{s_1, s_2, \ldots, s_T\}$ similar to the set $\mathcal{Y}_T$. $\mathcal{S}_t$ is often called a path. 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

$$\begin{aligned}
\Pr[Y_t = y_t, S_t = s_t | \mathcal{Y}_{t-1}, \mathcal{S}_{t-1}; \boldsymbol{\theta}] &= \Pr[Y_t = y_t | \mathcal{S}_t; \boldsymbol{\theta}] \Pr[S_t = s_t | \mathcal{S}_{t-1}; \boldsymbol{\theta}] \\
&= \begin{cases} f(y_t; \mu_0, \sigma_0^2) p_{00} & \text{if } s_t = 0, s_{t-1} = 0 \\ f(y_t; \mu_0, \sigma_0^2)(1 - p_{11}) & \text{if } s_t = 0, s_{t-1} = 1 \\ f(y_t; \mu_1, \sigma_1^2)(1 - p_{00}) & \text{if } s_t = 1, s_{t-1} = 0 \\ f(y_t; \mu_1, \sigma_1^2) p_{11} & \text{if } s_t = 1, s_{t-1} = 1 \end{cases} \\
&= \big(f(y_t; \mu_0, \sigma_0^2) p_{00}\big)^{(1-s_t)(1-s_{t-1})} \times \\
&\quad \big(f(y_t; \mu_0, \sigma_0^2)(1 - p_{11})\big)^{(1-s_t)s_{t-1}} \times \\
&\quad \big(f(y_t; \mu_1, \sigma_1^2)(1 - p_{00})\big)^{s_t(1-s_{t-1})} \times \\
&\quad \big(f(y_t; \mu_1, \sigma_1^2) p_{11}\big)^{s_t s_{t-1}}.
\end{aligned} \tag{15}$$

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 (15)

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

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 transition parameters p_{00} and p_{11} we find an expression with the parameter ζ,

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

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

$$\begin{aligned}\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) = &\sum_{t=1}^{T}\Big((1-s_t)\log f(y_t;\mu_0,\sigma_0^2) + s_t \log f(y_t;\mu_1,\sigma_1^2)\Big)+ \\ &\sum_{t=2}^{T}\Big((1-s_t)(1-s_{t-1})\log p_{00} + (1-s_t)s_{t-1}\log(1-p_{11})+ \\ &\quad s_t(1-s_{t-1})\log(1-p_{00}) + s_t s_{t-1}\log p_{11}\Big)+ \\ &\quad (1-s_1)\log\zeta + s_1\log(1-\zeta).\end{aligned} \tag{16}$$

The corresponding likelihood function is given by $\mathcal{L}_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) = \exp \ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta})$, and is much easier to evaluate than the true likelihood function in eq. (13). $\mathcal{L}_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta})$ applies to a specific path $\mathcal{S}_t$ whereas the true likelihood can be seen as a sum over all possible paths,

$$\mathcal{L}(\mathcal{Y}_T; \boldsymbol{\theta}) = \sum_{\mathcal{S}_T} \Pr(\mathcal{Y}_T|\mathcal{S}_T; \boldsymbol{\theta}) \Pr[\mathcal{S}_T; \boldsymbol{\rho}], \tag{17}$$

where $\Pr[\mathcal{S}_T; \boldsymbol{\rho}]$ gives the probability of a particular path. Though we prefer to optimize eq. (16), we cannot observe $\mathcal{S}_t$.

3.1.1 The Expectation and Maximization Steps

The EM-algorithm proposes to base the estimation on (16). Because we do not have actual observations on S_t, the EM-algorithm maximizes the expectation of the log likelihood function in (16) 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,

$$\mathrm{E}[S_t|\mathcal{Y}_T; \boldsymbol{\theta}] = \Pr[S_t = 0|\mathcal{Y}_T; \boldsymbol{\theta}] \cdot 0 + \Pr[S_t = 1|\mathcal{Y}_T; \boldsymbol{\theta}] \cdot 1 = \Pr[S_t = 1|\mathcal{Y}_T; \boldsymbol{\theta}]. \tag{18}$$

The last probability is a smoothed inference as in (12). Similarly, we find

$$\mathrm{E}[S_t S_{t-1}|\mathcal{Y}_T; \boldsymbol{\theta}] = \Pr[S_t = S_{t-1} = 1|\mathcal{Y}_T; \boldsymbol{\theta}]. \tag{19}$$

This approach would almost retain the attractive structure of the log likelihood function in (16). Almost, as the expectations of S_t and $S_t S_{t-1}$ depend on $\boldsymbol{\theta}$ and are calculated again via the recursion in (12). 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 (18) and (19) 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 $\boldsymbol{\theta}^{(k)}$

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathrm{E}\left[\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) | \mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}\right]. \tag{20}$$

Dempster et al. (1977) and Hamilton (1990) show that this sequence of $\boldsymbol{\theta}^{(k)}$ converges and produces a maximum of (14). As always, this maximum can be local, and may depend on starting values $\boldsymbol{\theta}^{(0)}$.

3.1.2 The Maximization Step in Detail

We now look at the maximization step in more detail. Our starting point is the likelihood function in (16), for which we calculate the expectation conditional on the data and parameters $\boldsymbol{\theta}^{(k-1)}$,

$$\ell_{\mathrm{EM}}\left(\mathcal{Y}_T; \boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}\right) = \mathrm{E}\left[\ell_{Y,S}(\mathcal{Y}_T, \mathcal{S}_T; \boldsymbol{\theta}) | \mathcal{Y}_T; \boldsymbol{\theta}^{(k-1)}\right] \tag{21}$$

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

$$\left.\frac{\partial \ell_{\mathrm{EM}}\left(\mathcal{Y}_T; \boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = 0. \tag{22}$$

Taking a closer look at (16), 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. In most Markov switching models, it is typically possible to evaluate the system of FOCs in those related to $\boldsymbol{\lambda}$ and those to $\boldsymbol{\rho}$.

We first look at differentiating (21) 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 that we find when we apply

the filter and smoother recursions in (9)–(12) with parameters $\boldsymbol{\theta}^{(k-1)}$. We find

$$\begin{aligned}\frac{\partial \ell_{\text{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial \mu_0} =& \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \log f(y_t;\mu_0,\sigma_0^2)}{\partial \mu_0} \\ =& \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(-\frac{1}{2}\log 2\pi - \log\sigma_0 - \frac{1}{2}\frac{(y_t-\mu_0)^2}{\sigma_0^2}\right)}{\partial \mu_0} \\ =& \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \frac{(y_t-\mu_0)}{\sigma_0^2}\end{aligned} \tag{23}$$

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

$$\mu_0^{(k)} = \frac{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} y_t}{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)}}. \tag{24}$$

This estimate for μ_0 can be interpreted as a weighted average of the observations, where the smoothed inferences 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

$$\begin{aligned}\frac{\partial \ell_{\text{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial \sigma_0} =& \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \log f(y_t;\mu_0,\sigma_0^2)}{\partial \sigma_0} \\ =& \frac{\partial \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(-\frac{1}{2}\log 2\pi - \log\sigma_0 - \frac{1}{2}\frac{(y_t-\mu_0)^2}{\sigma_0^2}\right)}{\partial \sigma_0} \\ =& \sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(\frac{(y_t-\mu_0)^2}{\sigma_0^3} - \frac{1}{\sigma_0}\right).\end{aligned} \tag{25}$$

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

$$\sigma_0^{(k)} = \sqrt{\frac{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)} \left(y_t - \mu_0^{(k)}\right)^2}{\sum_{t=1}^{T} \xi_{t|T,0}^{(k-1)}}}, \tag{26}$$

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

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

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 (12),

$$\tilde{p}_{ij,t+1}\equiv\Pr[S_{t+1}=i,S_t=j|\mathcal{Y}_T;\boldsymbol{\theta}^{(k-1)}]=\xi_{t|t,j}\cdot\frac{\xi_{t+1|T,i}}{\xi_{t+1|t,i}}p_{ij}^{(k-1)} \tag{27}$$

The derivative for p_{00} is given by

$$\begin{aligned}
\frac{\partial\ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial p_{00}}=&\frac{\partial\sum_{t=2}^{T}\tilde{p}_{00,t}\log p_{00}+\tilde{p}_{10,t}\log(1-p_{00})}{\partial p_{00}} \\
=&\sum_{t=2}^{T}\left(\frac{\tilde{p}_{00,t}}{p_{00}}-\frac{\tilde{p}_{10,t}}{1-p_{00}}\right).
\end{aligned} \tag{28}$$

Setting this expression to zero implies

$$p_{00}^{(k)}=\frac{\sum_{t=2}^{T}\tilde{p}_{00,t}}{\sum_{t=2}^{T}(\tilde{p}_{00,t}+\tilde{p}_{10,t})}=\frac{\sum_{t=2}^{T}\tilde{p}_{00,t}}{\sum_{t=2}^{T}\xi_{t-1|T,0}}. \tag{29}$$

This can be generalized to

$$p_{ij}^{(k)}=\frac{\sum_{t=2}^{T}\tilde{p}_{ij,t}}{\sum_{t=2}^{T}\xi_{t-1|T,j}}, \tag{30}$$

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

$$\begin{aligned}
\frac{\partial\ell_{\mathrm{EM}}\big(\mathcal{Y}_T;\boldsymbol{\theta},\boldsymbol{\theta}^{(k-1)}\big)}{\partial\zeta}=&\frac{\partial\left(\xi_{1|T,0}\log\zeta+\xi_{1|T,1}\log(1-\zeta)\right)}{\partial\zeta} \\
=&\frac{\xi_{1|T,0}}{\zeta}-\frac{\xi_{1|T,1}}{1-\zeta}.
\end{aligned} \tag{31}$$

Setting this expression to zero we find

$$\zeta^{(k)}=\xi_{t|T,0}^{(k-1)}. \tag{32}$$

3.1.3 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 variance 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. You can prove that each maximization step in the EM-algorithm yields an improvement (see Hamilton, 1990). In other words, for each new set of parameters $\boldsymbol{\theta}^{(k)}$, the log likelihood function in (14) *must* 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.[2]

3. Each step in the EM-algorithm yields an improvement in the likelihood function, though not necessarily monotonically. You have to specify a stopping criterion, which is best formulated for the increase in likelihood falling below a threshold.

3.2 Specification testing

Next to the estimates themselves, we are also interested in their precision. We can use the variance-covariance matrix of the estimates and the resulting standard errors in specification testing. Because the EM-algorithm maximizes the likelihood function, the standard properties of maximum likelihood (ML) estimators apply, given that the necessary regularity conditions are satisfied.[3] One crucial assumption is that the true parameter vector $\boldsymbol{\theta}_0$ is not on the boundary of its domain. This can be an issue for the probability parameter for the initial state ζ, which is therefor best excluded from the calculation of standard errors. If p_{00} or p_{11} are estimated close to zero or one, it is best to use their logit transformation.[4]

Assuming the regularity conditions are satisfied, the ML estimator $\hat{\boldsymbol{\theta}}$ converges in distribution

[2] Numerical issues can cause small increases, typically when (smoothed) inferences are close to zero.

[3] See Krolzig (2013, Ch. 6.6.1) for a discussion.

[4] The logit transformation of $0 < p < 1$ is given by $\log p - \log(1-p)$.

to a normal distribution,

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \mathcal{I}_0^{-1}). \tag{33}$$

Here $\mathcal{I}_0 \equiv \lim_{T\to\infty} \mathcal{I}(\boldsymbol{\theta}_0)$ denotes the asymptotic information matrix evaluated at the true parameter vector $\boldsymbol{\theta}_0$, where $\mathcal{I}(\boldsymbol{\theta}_0)$ gives the information matrix

$$\mathcal{I}(\boldsymbol{\theta}_0) \equiv -\,\mathrm{E}\left[\frac{1}{T}\frac{\partial^2 \ell(\mathcal{Y}_T;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = \mathrm{E}\left[\frac{\partial \ell(\mathcal{Y}_T;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\frac{\partial \ell(\mathcal{Y}_T;\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right]. \tag{34}$$

The second equality holds when the model is correctly specified.

In line with its definition, there are two ways of estimating the information matrix that are asymptotically equivalent, but may give different results in finite samples. The first is to use the Hessian matrix of the log likelihood function in eq. (14),

$$\hat{\mathcal{I}}_H(\hat{\boldsymbol{\theta}}) \equiv -\frac{1}{T}\frac{\partial^2 \ell(\mathcal{Y}_T;\hat{\boldsymbol{\theta}}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}. \tag{35}$$

The second way is to use the gradient, and construct the matrix as

$$\hat{\mathcal{I}}_G(\hat{\boldsymbol{\theta}}) \equiv \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{g}_t(\hat{\boldsymbol{\theta}})\boldsymbol{g}_t(\hat{\boldsymbol{\theta}})', \tag{36}$$

where

$$\boldsymbol{g}_t(\boldsymbol{\theta}) \equiv \frac{\partial \ell_t(y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{37}$$

gives the conditional score, that is, the gradient of the conditional log likelihood $\ell_t(y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta}) \equiv \log \Pr[Y_t = y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta}]$. By construction, $\ell(\mathcal{Y}_T;\boldsymbol{\theta}) = \sum_{t=1}^{T} \ell_t(y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta})$. The second estimator is often called the outer product of the gradient. Though both methods work, the same path-dependence that complicated the direct optimization of eq. (14) is a hindrance here, too. However, Hamilton (1993) shows that we can calculate the OPG in an easier way, circumventing long sequences of derivatives.

We start our derivation by using the relation between ℓ and ℓ_t to write

$$\boldsymbol{g}_t(\boldsymbol{\theta}) = \frac{\partial \ell(\mathcal{Y}_t;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial \ell(\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}, \tag{38}$$

so the difference in the scores of the full log likelihood function evaluated up to t and $t-1$. As in eq. (17), we can write

$$\ell(\mathcal{Y}_t;\boldsymbol{\theta}) = \log \Pr[\mathcal{Y}_t|\boldsymbol{\theta}] = \log \sum_{\mathcal{S}_t} \Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}] \Pr[\mathcal{S}_t|\boldsymbol{\rho}], \tag{39}$$

where we use that conditional on the path $\mathcal{S}_t$, the likelihood of $\mathcal{Y}_t$ only depends on the parameter vector $\boldsymbol{\lambda}$, and the likelihood of the path itself only depends on the parameter vector $\boldsymbol{\rho}$. Differentiation with respect to $\boldsymbol{\lambda}$ yields

$$\begin{aligned}
\frac{\partial \ell(\mathcal{Y}_t;\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} &= \frac{1}{\Pr[\mathcal{Y}_t|\boldsymbol{\theta}]} \sum_{\mathcal{S}_t} \frac{\partial \Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}]}{\partial \boldsymbol{\lambda}} \Pr[\mathcal{S}_t|\boldsymbol{\rho}] \\
&= \sum_{\mathcal{S}_t} \frac{\partial \log \Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}]}{\partial \boldsymbol{\lambda}} \frac{\Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}] \Pr[\mathcal{S}_t|\boldsymbol{\rho}]}{\Pr[\mathcal{Y}_t|\boldsymbol{\theta}]} \\
&= \sum_{\mathcal{S}_t} \frac{\partial \log \Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}]}{\partial \boldsymbol{\lambda}} \Pr[\mathcal{S}_t|\mathcal{Y}_t,\boldsymbol{\theta}] \\
&= \sum_{\mathcal{S}_t} \sum_{\tau=1}^{t} \frac{\partial \log \Pr[y_\tau|s_\tau,\boldsymbol{\lambda}]}{\partial \boldsymbol{\lambda}} \Pr[\mathcal{S}_t|\mathcal{Y}_t,\boldsymbol{\theta}] \\
&= \sum_{\tau=1}^{t} \sum_{s_\tau=0}^{1} \frac{\partial \log \Pr[y_\tau|s_\tau,\boldsymbol{\lambda}]}{\partial \boldsymbol{\lambda}} \Pr[s_\tau|\mathcal{Y}_t,\boldsymbol{\theta}] \\
&= \sum_{\tau=1}^{t} \frac{\partial \log \boldsymbol{f}'_\tau}{\partial \boldsymbol{\lambda}} \boldsymbol{\xi}_{\tau|t}. \qquad (40)
\end{aligned}$$

In the second equality, we use that for a differentiable function $g(x) > 0$, $\partial_x g(x) = g(x)\partial_x \log g(x)$. In the third equality, we apply Bayes' rule. In the fourth equality we use that the partial derivative for a given path reduces to a summation over the observations. In the fifth equality, we gather the probability mass that corresponds with the derivative of y_τ for a particular state s_τ. Because the derivative depends only on s_τ, the mass of the part of the path before and after s_τ sums to 1, and only $\Pr[s_\tau|\mathcal{Y}_t,\boldsymbol{\theta}]$ remains. In the final equality, we write the summation over the different states as a multiplication of the transpose of the Jacobian matrix of the (elementwise) log of $\boldsymbol{f}_\tau$ with the vector of smoothed inferences $\boldsymbol{\xi}_{\tau|t}$. The series of smoothed inferences $\{\boldsymbol{\xi}_{\tau|t}\}_{\tau=1}^{t}$ can be constructed with the Kim filter starting with $\boldsymbol{\xi}_{t|t}$. The derivatives of the log of $\boldsymbol{f}_\tau$ are typically easy to construct. Combining these derivations yields

$$\frac{\partial \ell_t(y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}} = \frac{\partial \ell(\mathcal{Y}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}} - \frac{\partial \ell(\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}} = \frac{\partial \log \boldsymbol{f}'_t}{\partial \boldsymbol{\lambda}} \boldsymbol{\xi}_{t|t} + \sum_{\tau=1}^{t-1} \frac{\partial \log \boldsymbol{f}'_\tau}{\partial \boldsymbol{\lambda}} (\boldsymbol{\xi}_{\tau|t} - \boldsymbol{\xi}_{\tau|t-1}). \quad (41)$$

This expression consists of two terms, the first related to the direct effect of observation t on the likelihood, and the second related to how the inference about the path changes because of it.

For the partial derivatives with respect to $\boldsymbol{\rho}$ we use a similar approach, starting with

$$\Pr[\mathcal{S}_t;\boldsymbol{\rho}] = \prod_{\tau=1}^{t} \Pr[s_\tau|s_{\tau-1},\boldsymbol{\rho}], \tag{42}$$

with $\Pr[s_1|s_0;\boldsymbol{\rho}] = \Pr[s_1;\boldsymbol{\rho}]$ to ease notation. Now we find

$$\begin{aligned}
\frac{\partial \ell(\mathcal{Y}_t;\boldsymbol{\lambda})}{\partial \boldsymbol{\rho}} &= \frac{1}{\Pr[\mathcal{Y}_t|\boldsymbol{\theta}]} \sum_{\mathcal{S}_t} \Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}] \frac{\partial \Pr[\mathcal{S}_t;\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \\
&= \sum_{\mathcal{S}_t} \frac{\partial \log \Pr[\mathcal{S}_t;\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \frac{\Pr[\mathcal{Y}_t|\mathcal{S}_t,\boldsymbol{\lambda}] \Pr[\mathcal{S}_t;\boldsymbol{\rho}]}{\Pr[\mathcal{Y}_t|\boldsymbol{\theta}]} \\
&= \sum_{\mathcal{S}_t} \frac{\partial \log \Pr[\mathcal{S}_t;\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \Pr[\mathcal{S}_t|\mathcal{Y}_t,\boldsymbol{\theta}] \\
&= \sum_{s_1=0}^{1} \frac{\partial \log \Pr[s_1;\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \Pr[s_1|\mathcal{Y}_t,\boldsymbol{\theta}] + \\
&\quad \sum_{\tau=2}^{t} \sum_{s_\tau=0}^{1} \sum_{s_{\tau-1}=0}^{1} \frac{\partial \log \Pr[s_\tau|s_{\tau-1},\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \Pr[s_\tau, s_{\tau-1}|\mathcal{Y}_t,\boldsymbol{\theta}].
\end{aligned} \tag{43}$$

The derivatives for $\log \Pr[s_\tau|s_{\tau-1},\boldsymbol{\rho}]$ are easy to find, and the recursion in eq. (27) can be used to construct the smoothed inferences $\Pr[s_\tau, s_{\tau-1}|\mathcal{Y}_t,\boldsymbol{\theta}]$. Finally, we combine these derivations to

$$\begin{aligned}
&\frac{\partial \ell_t(y_t|\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial \boldsymbol{\rho}} = \frac{\partial \ell(\mathcal{Y}_t;\boldsymbol{\theta})}{\partial \boldsymbol{\rho}} - \frac{\partial \ell(\mathcal{Y}_{t-1};\boldsymbol{\theta})}{\partial \boldsymbol{\rho}} \\
&= \sum_{s_t=0}^{1} \sum_{s_{t-1}=0}^{1} \frac{\partial \log \Pr[s_t|s_{t-1},\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \Pr[s_t, s_{t-1}|\mathcal{Y}_t,\boldsymbol{\theta}] + \\
&\sum_{\tau=2}^{t-1} \sum_{s_\tau=0}^{1} \sum_{s_{\tau-1}=0}^{1} \frac{\partial \log \Pr[s_\tau|s_{\tau-1},\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} (\Pr[s_\tau, s_{\tau-1}|\mathcal{Y}_t,\boldsymbol{\theta}] - \Pr[s_\tau, s_{\tau-1}|\mathcal{Y}_{t-1},\boldsymbol{\theta}]) + \\
&\sum_{s_1=0}^{1} \frac{\partial \log \Pr[s_1;\boldsymbol{\rho}]}{\partial \boldsymbol{\rho}} \big(\Pr[s_1|\mathcal{Y}_t,\boldsymbol{\theta}] - \Pr[s_1|\mathcal{Y}_{t-1},\boldsymbol{\theta}]\big).
\end{aligned} \tag{44}$$

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. The data is available in the file `MSExample_MSCIUS.xls` on my website[5] and on Mendeley[6]. The returns are given in %.

[5] See `http://personal.eur.nl/kole`.

[6] See `https://data.mendeley.com/datasets/8yvfgdwnbr/1`.

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

$$\mu_0 = 0.04 \qquad \sigma_0 = 1 \qquad p_{11} = 0.80 \qquad \zeta = 0.50$$
$$\mu_1 = -0.04 \qquad \sigma_1 = 4 \qquad p_{22} = 0.80$$

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 inferences. The first forecast probabilities are given by ζ and $1 - \zeta$. Based on the first return of -1.01923, the inferences 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 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 (inferences for regime 2 exceed 0.5). Consequently, when we smooth the series of inferences, 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 (14) 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. We also calculate the standard errors for the estimates, using a numerical approximation to the Hessian and the outer product of the gradient. Because $\hat{\zeta} = 1$, which is on the boundary of the support, we do not calculate a standard error for it. The standard errors of both methods are close for the mean and transition parameters, but not for the volatility parameters, which may point at misspecification.

In Table 3 we report the forecast, inference and smoothed inferences for the first ten returns, based on the parameters estimates produced by the EM-algorithm. Compared to Table 1, we

Table 1: Inferences for the first ten returns.

		forecast probabilities		inference probabilities		smoothed inf. probabilities	
observation	return	$S_t = 0$	$S_t = 1$	$S_t = 0$	$S_t = 1$	$S_t = 0$	$S_t = 1$
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, inferences and smoothed inferences. 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 $\zeta = 0.50$.

Table 2: Steps of the EM-algorithm

	starting values	iteration 1	iteration 2	iteration 3	optimal solution	standard errors Hessian	standard errors OPG
μ_0	0.0400	0.1426	0.1980	0.2240	0.1573	0.0526	0.0531
σ_0	1.0000	1.1445	1.2182	1.2645	1.5594	0.0526	0.0171
μ_1	−0.0400	−0.1262	−0.1887	−0.2324	−0.2988	0.1725	0.1746
σ_1	4.0000	3.1417	3.0916	3.1030	3.4068	0.1639	0.0235
p_{11}	0.8000	0.8222	0.8345	0.8532	0.9770	0.0069	0.0066
p_{22}	0.8000	0.7899	0.8072	0.8195	0.9484	0.0175	0.0147
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 (14). Standard errors are calculated using the numerical approximation of the Hessian and the outer product of the gradient (OPG).

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

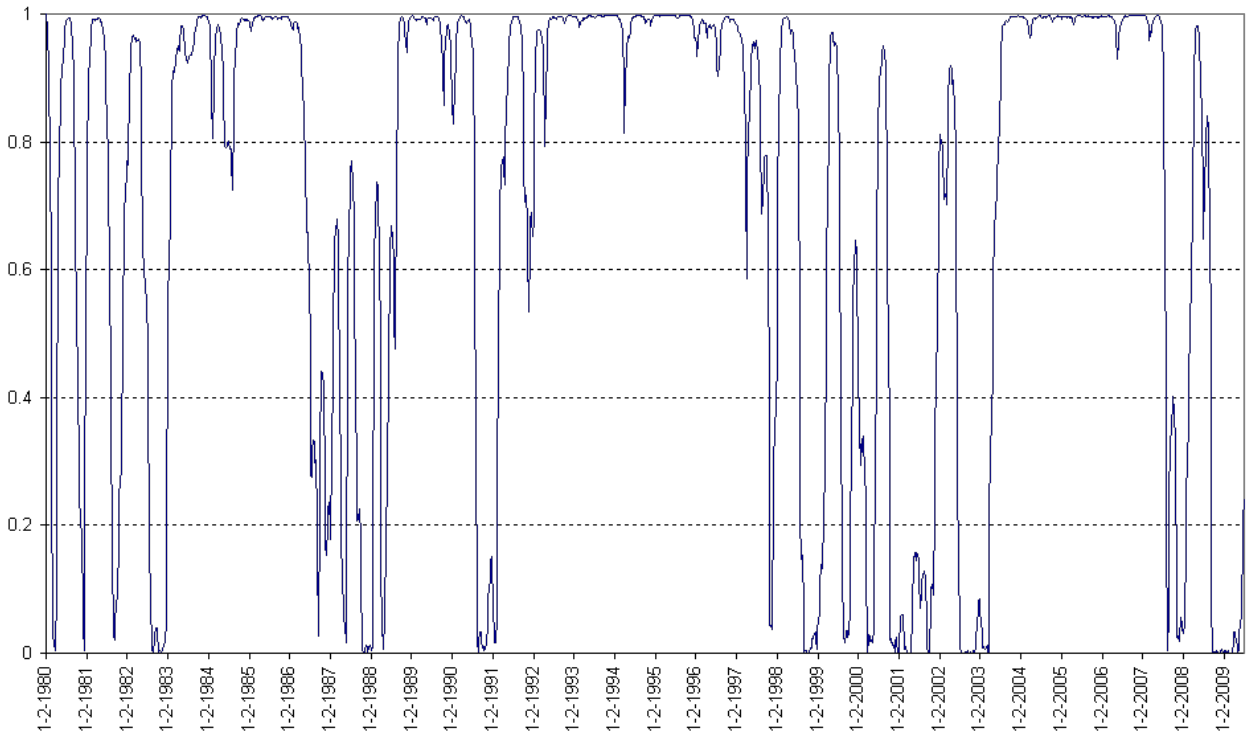
		forecast probabilities		inference probabilities		smoothed inf. probabilities	
observation	return	$S_t = 0$	$S_t = 1$	$S_t = 0$	$S_t = 1$	$S_t = 0$	$S_t = 1$
1	−1.01923	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
2	2.64830	0.97697	0.02303	0.97411	0.02589	0.97756	0.02244
3	1.54639	0.95301	0.04699	0.97184	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, inferences and smoothed inferences. 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.

see the regimes are better defined now: the probabilities are either close to zero or to one. The inferences 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 inferences 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.

Figure 1: Smoothed Inference Probability for Regime 0



This figure shows the smoothed inferences for regime 0 over time for the US stock market. The probabilities are constructed using the filter recursion in (9) and (10) and the smoother recursion of Kim (1994) in (12). 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.

Hamilton, J. D. (1993). Estimation, inference and forecasting of time series subject to changes in regime. In Maddala, G., Rao, C., and Vinod, H., editors, *Handbook of Statistics*, volume 11, chapter 9, pages 231–260. Elsevier Science Publishers B.V.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ, USA.

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

Kole, E. and Dijk, D. (2017). How to identify and forecast bull and bear markets? *Journal of Applied Econometrics*, 32(1):120–139.

Krolzig, H.-M. (2013). *Markov-switching vector autoregressions: Modelling, statistical inference, and application to business cycle analysis*, volume 454 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag Berlin Heidelberg.