Anders Rahbek
Rasmus Søndergaard Pedersen
University of Copenhagen

October 2021

Part VII
Estimation of Finite State SV Models

In this chapter we discuss how to obtain likelihood-based estimators in SV models, where the switching driven by a finite-state first-order Markov chain, such as the two-state SV model:

Example VII.1 *The two-state SV model is given by,*

$$x_t = \sigma_t z_t,$$

for $t = 1, 2, ..., T$ and where the innovations z_t are i.i.d.$N(0,1)$ and independent of the unobserved volatility process (σ_t). The volatility is given by the unobserved switching process (the Markov chain) $s_t \in \{1, 2\}$, such that

$$\sigma_t = \begin{cases} h_1 & \text{if } s_t = 1 \\ h_2 & \text{if } s_t = 2 \end{cases},$$

and where the switching between the two states is governed by the transition matrix P given by

$$P = \begin{pmatrix} p(s_t = 1|s_{t-1} = 1) & p(s_t = 1|s_{t-1} = 2) \\ p(s_t = 2|s_{t-1} = 1) & p(s_t = 2|s_{t-1} = 2) \end{pmatrix} = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11} & 1 - p_{22} \\ 1 - p_{11} & p_{22} \end{pmatrix}.$$

Hence the parameters we wish to estimate are

$$\theta = \{h_1, h_2, p_{11}, p_{22}\},$$

with $h_1, h_2 \in (0, \infty)$ and $p_{11}, p_{22} \in (0, 1)$. Much like in the GARCH model where the initial value σ_0^2 is unknown, the initial value s_0 (and hence σ_0^2) is not known.

Example VII.2 *For the general N-state case where $x_t = \sigma_t z_t$, the unobserved volatility sequence can take N values $h_1, ..., h_N$ respectively, according*

to the finite state switching process s_t on $\{1, ..., N\}$. The distribution of the one-step transition of the Markov chain s_t is given by the transition matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{N1} \\ \vdots & \ddots & \vdots \\ p_{1N} & \cdots & p_{NN} \end{pmatrix},$$

where $p_{ij} = P(s_t = j \mid s_{t-1} = i)$. The parameters to be estimated are therefore $\theta = (h_1, ..., h_N, p_{11}, ..., p_{NN})$, where for $(p_{11}, ..., p_{NN})'$ we note the restriction that each column of the transition matrix must sum to one, that is for example $1 = p_{11} + ... + p_{1N}$.

Even though the two-state model is simple, it is widely applied, and moreover, as we shall see, somewhat involved to do estimation in. By definition the two-state SV model is a special case of the general finite (N-)state SV model and we shall formulate the results in a way such that also this model can be handled. In fact, the finite state SV models can also be viewed as examples of the wide class of so-called Hidden Markov Models (HMM) and/or Markov Switching (MS) models, which are widely applied – also in the field of macroeconometrics and microeconometrics, see also Section VII.5 below.

In terms of estimation, we note that writing down the log-likelihood function for the observations,

$$(x_1, ..., x_T),$$

is slightly involved, due to the fact that the path of $(s_t : t = 1, \ldots, T)$ is unobserved to the econometrician, which we will discuss in the following section. Further references for general MS models include Hamilton (1994, ch.22), who discuss their use in macro time series, MacDonald and Zucchini (1997), who provide a general statistical introduction, and Lange and Rahbek (2008), who focus on financial time series.

VII.1 Maximum likelihood estimation (MLE)

First, we make some initial considerations about the likelihood function and introduce some useful notation. Recall that we observe (the log-returns),

$$X_{0:T} = (x_0, ..., x_T),$$

while we do not observe the volatility and the switching variables,

$$S_{0:T} = (s_0, ..., s_T).$$

VII.1.1 A side note: Treating s_t as observed

If the switching variables, or the states, $S_{0:T}$ were observed the likelihood function could be computed as follows:

First note that as in autoregressive models, we can factorize the density conditional on (x_0, s_0) as follows (using $f(\cdot)$ to denote generic densities),

$$\begin{aligned} f(X_{1:T}, S_{1:T}|x_0, s_0) &= f(x_1, s_1, ..., s_T, x_T|x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t, s_t|x_{t-1}, s_{t-1}, ..., x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t|s_t) f(s_t|x_{t-1}, s_{t-1}, ..., x_0, s_0) \\ &= \prod_{t=1}^{T} f(x_t|s_t) p_{s_{t-1}s_t}, \end{aligned} \tag{VII.1}$$

where we have used that s_t depends only on s_{t-1} and not the remaining past values. Here, as z_t are $i.i.d.\mathrm{N}(0,1)$, we have

$$f(x_t|s_t) = \frac{1}{\sqrt{2\pi h_{s_t}^2}} \exp\left(-\frac{x_t^2}{2h_{s_t}^2}\right),$$

while to emphasize that s_t is characterized by probabilities rather than a density, we apply the notation $p_{s_{t-1}s_t}$ from the definition of the transition matrix P, that is

$$p_{s_{t-1}s_t} = p_{ij} \quad \text{if } (s_{t-1}, s_t) = (i, j).$$

Likewise, we use the notation

$$h_{s_t} = \begin{cases} h_1 & \text{if } s_t = 1 \\ h_2 & \text{if } s_t = 2 \end{cases}.$$

To get the full density, which is not conditional on (x_0, s_0), we can write,

$$f(X_{0:T}, S_{0:T}) = f(x_1, s_1, \ldots, , x_T, s_T|x_0, s_0) f(x_0, s_0).$$

Now in the two-state case s_1 can be either 1 or 2. Recall that the invariant probabilities were computed in Part VI as

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} p(s_1 = 1) \\ p(s_1 = 2) \end{pmatrix} = \begin{pmatrix} \frac{1}{2-p_{11}-p_{22}}(1-p_{22}) \\ \frac{1}{2-p_{11}-p_{22}}(1-p_{11}) \end{pmatrix}, \tag{VII.2}$$

and we could therefore write

$$f(x_0, s_0) = f(x_0|s_0)\, v_{s_0}.$$

This way we have the full likelihood function in terms of $\theta = (h_1, h_2, p_{11}, p_{22})'$. However, it is simpler to set $v := (v_1, 1 - v_1)$ and extend θ to include also v_1, or even to set for example $v_1 = \frac{1}{2}$ (that is, fixed and known). With large samples these choices matter very little.

Likewise, for the N-state case where $v = (v_1, ..., v_N)'$ one often instead of using the invariant probabilities expressed in terms of parameters in θ, choose to just let the parameters to be estimated to be (θ, v), or alternatively fix v as for example $v = \left(\frac{1}{N}, ..., \frac{1}{N}\right)'$ as in the two-state case. The latter is obviously the easiest.

In any case, collecting all terms we get the full log-likelihood function to be,

$$L_T(X_{0:T}, S_{0:T}; \theta) = \log\left(v_{s_0} f(x_0|s_0) \prod_{t=1}^{T} p_{s_{t-1}s_t} f(x_t \mid s_t)\right), \qquad \text{(VII.3)}$$

where $v = (v_1, ..., v_N)'$.[1]

VII.1.2 Treating s_t as unobserved

As s_t is not observed, the likelihood function must be evaluated based solely on the sequence $(x_1, ..., x_T)$. The direct way to do this is by simply summing (VII.3) over all possible paths for the state process s_t, that is

$$L_T(X_{0:T}; \theta) = \sum_{(s_0,...,s_T)\in\{1,...,N\}^{T+1}} L_T(X_{0:T}, S_{0:T}; \theta).$$

Unfortunately, this formulation leads to an exponentially growing number of terms, N^{T+1}, and is therefore infeasible even for moderate sample sizes. One could actually overcome this and making it grow linearly in T by formulating the likelihood function in terms of matrix products. However, the likelihood function is non-linear in θ and there are complex restrictions limiting the parameter space (primarily coming from the restrictions on the parameters

[1] Alternatively, one could consider the conditional log-likelihood based on (VII.1),

$$L_{T,c}(x_0, ..., x_T, s_0, ..., s_T; \theta) = \log\left(\prod_{t=1}^{T} p_{s_{t-1}s_t} f(x_t \mid s_t)\right).$$

in the transition matrix P), which means that we will take the alternative route of applying a so-called filtering algorithm as is quite standard in the literature, and which avoids these problems.

For the case of $N = 2$ states, we note that the density of $x_t|X_{0:t-1}$ is a weighted mixture of the two possible states,

$$\begin{aligned} f\left(x_t|X_{0:t-1}\right) &= f\left(x_t, s_t = 1|X_{0:t-1}\right) + f\left(x_t, s_t = 2|X_{0:t-1}\right) \\ &= f\left(x_t|s_t = 1, X_{0:t-1}\right) P\left(s_t = 1|X_{0:t-1}\right) \\ &\quad + f\left(x_t|s_t = 2, X_{0:t-1}\right) P\left(s_t = 2|X_{0:t-1}\right) \\ &= f\left(x_t|s_t = 1\right) P\left(s_t = 1|X_{0:t-1}\right) \\ &\quad + f\left(x_t|s_t = 2\right) P\left(s_t = 2|X_{0:t-1}\right) \end{aligned}$$

with weights given by the predicted probabilities, $P\left(s_t = i|X_{0:t-1}\right)$, $i = 1, 2$. By construction, the predicted probability, $P\left(s_t = i|X_{0:t-1}\right)$, is the best guess of the probability of regime i at time t given the observations until time $t-1$, $X_{0:t-1}$. The log-likelihood function (conditional on some known x_0) is hence given by

$$\begin{aligned} L_T(x_1, ..., x_T; \theta) &= \sum_{t=1}^{T} \log f\left(x_t|X_{0:t-1}; \theta\right) \\ &= \sum_{t=1}^{T} \log \left(\sum_{i=1}^{2} f\left(x_t|s_t = i; \theta\right) P\left(s_t = i|X_{0:t-1}; \theta\right) \right). \end{aligned} \tag{VII.4}$$

Note that

$$f\left(x_t|s_t = i; \theta\right) = \frac{1}{\sqrt{2\pi h_i^2}} \exp\left(-\frac{x_t^2}{2h_i^2}\right) \tag{VII.5}$$

is straightforward to evaluate given θ, and the challenging part of the estimation is to evaluate the predicted probabilities $P\left(s_t = i|X_{0:t-1}; \theta\right)$. We will consider an algorithm that evaluates the probabilities recursively.

VII.1.3 Filtering Algorithm

Suppose that we already have some values of the conditional regime probabilities at time $t-1$, $P\left(s_{t-1} = i|X_{0:t-1}\right)$, which is typically labelled the *filtered*

probability. Then the predicted probabilities are simply given by

$$\begin{aligned} P\left(s_t = i | X_{0:t-1}; \theta\right) &= P\left(s_t = i, s_{t-1} = 1 | X_{0:t-1}; \theta\right) + P\left(s_t = i, s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= P\left(s_t = i | s_{t-1} = 1, X_{0:t-1}; \theta\right) P(s_{t-1} = 1 | X_{0:t-1}; \theta) \\ &\quad + P\left(s_t = i | s_{t-1} = 2, X_{0:t-1}; \theta\right) P\left(s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= P\left(s_t = i | s_{t-1} = 1; \theta\right) P(s_{t-1} = 1 | X_{0:t-1}; \theta) \\ &\quad + P\left(s_t = i | s_{t-1} = 2; \theta\right) P\left(s_{t-1} = 2 | X_{0:t-1}; \theta\right) \\ &= \sum_{j=1}^{2} p_{ji} P\left(s_{t-1} = j | X_{0:t-1}; \theta\right). \quad \textbf{(prediction step)} \end{aligned} \tag{VII.6}$$

In order to make the algorithm recursive, we need a way to filtered probability at time t, $P\left(s_t = i | X_{0:t}; \theta\right)$, based on the predicted probability $P\left(s_t = i | X_{0:t-1}; \theta\right)$. To do so, we make use of the definition of conditional probability:

$$P\left(s_t = i | X_{0:t}; \theta\right) = \frac{f\left(s_t = i, x_t | X_{0:t-1}; \theta\right)}{f\left(x_t | X_{0:t-1}; \theta\right)}.$$

The numerator, given by the joint conditional density of x_t and $\{s_t = i\}$, equals

$$\begin{aligned} f\left(s_t = i, x_t | X_{0:t-1}; \theta\right) &= f\left(x_t | s_t = i, X_{0:t-1}; \theta\right) P\left(s_t = i | X_{0:t-1}; \theta\right) \\ &= f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{0:t-1}; \theta\right), \end{aligned}$$

and hence is a product of the conditional density in (VII.5) and predicted probability in (VII.6). Likewise, the denominator is the likelihood contribution in (VII.4). Hence, we have that the filtered probability is

$$P\left(s_t = i | X_{1:t}; \theta\right) = \frac{f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{1:t-1}; \theta\right)}{\sum_{i=1}^{2} f\left(x_t | s_t = i; \theta\right) P\left(s_t = i | X_{1:t-1}; \theta\right)}, \quad \textbf{(filtering step)}$$

and we emphasize that all terms are straightforward to evaluate. Given an initial value $P\left(s_0 = i | x_0; \theta\right)$ for $i = 1, 2$, e.g. $P\left(s_0 = i | x_0; \theta\right) = P(s_0 = i)$, we note that all predicted and filtered probabilities at times $t = 1, \ldots, T$ are determined recursively by the outlined steps, which are typically referred to as the *Hamilton algorithm*.

VII.1.4 Asymptotic Distribution and Inference

Asymptotic theory for the maximum likelihood estimators in HMMs is an active area of research, see, e.g., Kasahara and Shimotsu (2019) for a recent overview and the references herein. Proving asymptotic normality of

the maximum likelihood estimator, $\hat{\theta}_T$, in HHMs is typically much more demanding compared to the proofs for the GARCH-type models considered in Parts I-III. However, there exist high level conditions for asymptotic normality implying that

$$\sqrt{T}\left(\hat{\theta}_T - \theta_0\right)$$

has the usual asymptotic distribution in terms of the inverse of the information, see, e.g., Bickel et al. (1999) and Douc et al. (2004). The conditions are essentially that s_t is weakly mixing for our N-state SV models, and that θ_0 is an interior point of the parameter space.

An empirically important hypothesis to test is how many regimes are needed in order to characterize a given data series, e.g. if only $N = 1$ regime is needed instead of $N = 2$ regimes. Testing such a hypothesis is complicated due to the fact that parameter(s) of state 2, say the volatility h_2, is not identified under the null hypothesis of only one state. This violates the standard conditions of likelihood ratio testing, and the LR statistic does not have a standard χ^2 distribution. Several non-standard testing methods have been suggested in order to circumvent this issue, see e.g. Davies (1987), Andrews and Ploberger (1994), Hansen (1996), Cho and White (2007), and Carrasco et al. (2014).

VII.2 Predicted, Filtered, and Smoothed probabilities

Note that a by-product of the filtering algorithm, that evaluates the log-likelihood function, is that it produces the predicted probability $P\left(s_t = i|X_{0:t-1};\theta\right)$, i.e. the probability of being in state i in the next period. Such a probability is typically interesting for prediction purposes. As an example, one may quantify the probability of being in a high volatility state tomorrow given the information available today. Predicted probabilities extend to longer horizons, i.e. $P\left(s_{t+k} = i|X_{0:t};\theta\right)$ for any $k > 1$. Likewise, the filtered probability, $P\left(s_t = i|X_{0:t};\theta\right)$, may be used for addressing the probability of being in a high volatility regime today taking into account today's return y_t. This is sometimes referred to as real-time regime classification.

Lastly, it might be of interest to characterize the regime probabilities for a period given the information up to today, i.e. to compute the so-called smoothed probabilities $P\left(s_t = i|X_{0:T};\theta\right)$, where we note that the conditioning is based on all observations $X_{0:T} = (x_0, \ldots, x_T)$. Whereas the predicted and filtered probabilities are given recursively forward in time, the smoothed

probabilities are given by combining forward and backward recursions, as outlined in the following.

For $t = 1, ..., T$ define for $j = 1, 2$ the densities

$$\begin{aligned} a_t(j) &= f(x_0, ..., x_t, s_t = j; \theta), \\ b_t(j) &= f(x_{t+1}, ..., x_T \mid s_t = j; \theta), \end{aligned}$$

such that $a_t(j)$ denotes the joint density of the first t observations and of s_t. Similarly, $b_t(j)$ denotes the density of $x_{t+1}, \ldots, x_T$ given $s_t = j$. In terms of a_t and b_t, the smoothed probabilities $P(s_t = i|X_{0:T}; \theta)$ for $i = 1, 2$ can be computed as

$$\begin{aligned} &P(s_t = i|X_{0:T}; \theta) \\ &= \frac{f(x_0, ..., x_T, s_t = i; \theta)}{f(x_0, ..., x_T; \theta)} \\ &= \frac{f(x_0, ..., x_T, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_0, ..., x_T\ , s_t = j; \theta)} \\ &= \frac{f(x_{t+1}, ..., x_T|x_0, ..., x_t, s_t = i; \theta) f(x_0, ..., x_t, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_{t+1}, ..., x_T|x_1, ..., x_t, s_t = j; \theta) f(x_0, ..., x_t, s_t = j; \theta)} \\ &= \frac{f(x_{t+1}, ..., x_T|s_t = i; \theta) f(x_0, ..., x_t, s_t = i; \theta)}{\sum_{j=1}^{2} f(x_{t+1}, ..., x_T|s_t = j; \theta) f(x_0, ..., x_t, s_t = j; \theta)} \\ &= \frac{b_t(i) a_t(i)}{\sum_{j=1}^{2} b_t(j) a_t(j)}. \end{aligned} \tag{VII.7}$$

Note that we in particular used that

$$f(x_{t+1}, ..., x_T|x_1, ..., x_t, s_t = i) = f(x_{t+1}, ..., x_T|s_t = i),$$

which follows by the definition of the SV model with s_t the underlying Markov chain.

The sequence $a_t(\cdot)$ can be computed recursively using the *forward* algorithm given by

$$\begin{aligned} a_t(j) &= f(s_t = j, x_1, ..., x_t; \theta) \\ &= \sum_{i=1}^{2} f(s_{t-1} = i, s_t = j, x_1, ..., x_t) \\ &= \sum_{i=1}^{2} f(x_t|s_t = j; \theta) p_{ij} a_{t-1}(i). \end{aligned} \tag{VII.8}$$

The algorithm is initiated by putting $a_0(j) = f(x_0|s_0 = j;\theta)v_j$ for some fixed distribution $v = (v_1, v_2)$ on $\{1, 2\}$ for example $v = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Finally the sequence $b_t(\cdot)$ can be computed using the *backward* algorithm given by

$$\begin{aligned} b_t(i) &= f(x_{t+1}, ..., x_T|s_t = i;\theta) \\ &= \sum_{j=1}^{2} f(s_{t+1} = j, x_{t+1}, ..., x_T|s_t = i;\theta) \\ &= \sum_{j=1}^{2} f\left(x_{t+1}|s_{t+1} = j;\theta\right) f\left(x_{t+2}, ..., x_T|s_{t+1} = j;\theta\right) f\left(s_{t+1} = j|s_t = i;\theta\right) \\ &= \sum_{j=1}^{2} f(x_{t+1}|s_{t+1} = j;\theta) b_{t+1}(j) p_{ij}. \end{aligned} \tag{VII.9}$$

The algorithm can be initiated by putting $b_T(j) = 1$ for all $j = 1, 2$.

VII.3 The General Case

In the following we give a brief outline of recursions in relation to the general N-state model introduced in Example VII.2. Most of the quantities are written on vector form, which is convenient for implementation in a programming language. For $s_t \in \{1, \ldots, N\}$, define the N-dimensional vector of conditional densities,

$$\eta_{t,\theta} = \begin{pmatrix} f(x_t|s_t = 1, X_{0:t-1};\theta) \\ \vdots \\ f(x_t|s_t = N, X_{0:t-1};\theta) \end{pmatrix},$$

and the predicted and filtered probabilities.

$$\xi_{t|t-1,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:t-1};\theta) \\ \vdots \\ P(s_t = N|X_{0:t-1};\theta) \end{pmatrix} \quad \text{and} \quad \xi_{t|t,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:t};\theta) \\ \vdots \\ P(s_t = N|X_{0:t};\theta) \end{pmatrix}.$$

The log-likelihood function for the observations $x_1, \ldots, x_T$ is then given by

$$\begin{aligned} L_T(\theta) &= \sum_{t=1}^{T} \log\left(\sum_{i=1}^{N} f\left(x_t|s_t = i;\theta\right) P\left(s_t = i|X_{0:t-1};\theta\right)\right) \\ &= \sum_{t=1}^{T} \log\left(\eta'_{t,\theta}\xi_{t|t-1,\theta}\right). \end{aligned}$$

Moreover, the filtering algorithm for $t = 1, \ldots, T$ is

$$\xi_{t|t,\theta} = \frac{1}{\eta'_{t,\theta}\xi_{t|t-1,\theta}} \left(\eta_{t,\theta} \odot \xi_{t|t-1,\theta}\right) \quad \text{(Filtering step)},$$
$$\xi_{t+1|t,\theta} = P\xi_{t|t,\theta} \quad \text{(Prediction step)},$$

where $\odot$ denotes element-by-element (so-called Hadamard) multiplication. The recursions may be initiated at some given probabilities such as the uniform $\xi_{1|0,\theta} = (N^{-1}, \ldots, N^{-1})'$ or at the stationary probabilities $\xi_{1|0,\theta} = v$, with v defined in (VI.10) in Part VI.

Moreover, it is straightforward to show that the k-step predicted probabilities are given by

$$\xi_{t+k|t,\theta} = \begin{pmatrix} P(s_{t+k} = 1|X_{0:t};\theta) \\ \vdots \\ P(s_{t+k} = N|X_{0:t};\theta) \end{pmatrix} = P^k \xi_{t|t,\theta},$$

and, lastly, the smoothed probabilities are given backwards by

$$\xi_{t|T,\theta} = \begin{pmatrix} P(s_t = 1|X_{0:T};\theta) \\ \vdots \\ P(s_t = N|X_{0:T};\theta) \end{pmatrix} = \xi_{t|t,\theta} \odot \left(P' \left[\xi_{t+1|T,\theta} \oslash \xi_{t+1|t,\theta}\right]\right),$$

where $\oslash$ denotes element-by-element division (see e.g. Hamilton, 1994, ch.22). The backward recursions may be initiated at the time T filtered probabilities, $\xi_{T|T,\theta}$.

VII.4 An Alternative Estimation Approach

An alternative to maximizing the log-likelihood function is the so-called Expectation-Maximization (EM) algorithm. The main idea is that if the latent process, $(s_1, \ldots, s_T)$, was observed, then it would be straightforward to maximize the joint log-likelihood function, as derived in (VII.3). The EM algorithm relies on maximizing the expected log-likelihood with the expectation taken conditionally over the observed sample $(x_1, \ldots, x_T)$, and this can be shown to yield the MLE. The expected log-likelihood function is derived via an algorithm similar to the one used for obtaining the aforementioned smoothed probabilities. This is known as the expectation step (E-step) of the algorithm.

VII.5 Extensions

The MS SV model can be extended in many directions. First, the classical AR and ARCH models considered in the first chapters may have regime switching parameters. Specifically, we may consider an MS-AR model with stochastic volatility where for $x_t \in \mathbb{R}$,

$$x_t = \mu_{s_t} + \rho_{s_t} x_{t-1} + \sigma_{s_t} z_t, \quad z_t \ i.i.d. N(0,1),$$

with $(\mu_{s_t}, \rho_{s_t}, \sigma_{s_t}) = (\mu_i, \rho_i, h_i) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$ when $s_t = i$, $i = 1, \ldots, N$, and (s_t) an N-state Markov chain, and where (s_t) and (z_t) are independent. Obviously, we could let z_t have another distribution, e.g. a Student's t-distribution (potentially with regime-switching degrees of freedom, in order to allow for time-varying tail heaviness). Likewise, we may as in Cai (1994) consider an MS-ARCH model, with

$$\begin{aligned} x_t &= \sigma_t z_t, \quad z_t \ i.i.d. N(0,1) \\ \sigma_t^2 &= \omega_{s_t} + \alpha_{s_t} x_{t-1}^2, \end{aligned}$$

with $(\omega_{s_t}, \alpha_{s_t}) = (\omega_i, \alpha_i) \in (0, \infty) \times [0, \infty)$ when $s_t = i$, $i = 1, \ldots, N$, and with (s_t) and (z_t) independent .

We may also consider multivariate extensions, such as MS-VAR models with stochastic covolatility, where for $x_t \in \mathbb{R}^d$

$$x_t = \mu_{s_t} + A_{s_t} x_{t-1} + \Sigma_{s_t}^{1/2} z_t, \quad z_t \ i.i.d. N(0, I_d),$$

with μ_{s_t} and A_{s_t} are respectively $(d \times 1)$ and $(d \times d)$ matrices and $\Sigma_{s_t}^{1/2}$ is the matrix square-root of a $(d \times d)$ positive definite matrix Σ_{s_t} – all allowed to be regime switching. Such a model has been widely used in empirical macro and finance, such as in Guidolin and Timmermann (2004) who apply the model to detect "bull" and "bear" states in UK stock and bond markets, and to construct optimal portfolio allocation, taking into account potential switching in economic regimes, see also Barberis (2000). Note that all of these extensions are, by definition, not SV models. Estimation in such models are more or less straightforward, and the filtering algorithms remains (up to some additional conditioning) essentially unchanged.

Lastly, one may allow for (strictly) exogenous covariates in the models and allow the transition probabilities to be time-varying, e.g. by letting the state probabilities depend on lagged observations, $x_1, \ldots, x_{t-1}$, or covariates; see e.g. Bec et al. (2008).

References

Andrews, D.W.K. and W. Ploberger, 1994, "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative", *Econometrica*, 62: 1383–1414.

Barberis, N., 2000, "Investing for the Long Run when Returns Are Predictable", *The Journal of Finance*, 55, 225–264.

Bec, F., A. Rahbek, and N. Shephard, 2008, "The ACR Model: A Multivariate Dynamic Mixture Autoregression", *Oxford Bulletin of Economics and Statistics*, 70, 583–618.

Bickel, P.J., Y. Ritov, and T. Rydén, 1998, "Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models", *Annals of Statistics*, 26, 1614–1635.

Cai, J., 1994, "A Markov Model of Switching-Regime ARCH", *Journal of Business & Economic Statistics*, 12, 309–316.

Carrasco, M., L. Hu, and W. Ploberger, 2014, "Optimal Test for Markov Switching Parameters, *Econometrica*, 82, 765–784.

Cho, J.S. and H. White, 2007, "Testing for Regime Switching", *Econometrica*, 75m 1671–1720.

Davies, R.B., 1987, "Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative", *Biometrika*, 74, 33–43.

Douc R., É. Moulines, and T. Rydén, 2004, "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime", *Annals of Statistics*, 32, 2254–2304.

Guidolin, M. and A. Timmermann, 2004, "Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns", *The Economic Journal*, 115, 111–143.

Hamilton, J.D., 1994, *Time Series Analysis*, Princeton University Press.

Hansen, B.E., 1996, "Inference when a Nuisance Parameter is Not Identified Under the Null Hypothesis", *Econometrica*, 64, 413–430.

Kasahara, H. and K. Shimotsu, 2019, "Asymptotic Properties of the Maximum Likelihood Estimator in Regime Switching Econometric Models", *Journal of Econometrics*, 208, 442–467.

Lange T. and A. Rahbek, 2008, "Regime Switching Time Series Models: A Survey", *Handbook of Financial Time Series*.

MacDonald, I.L. and W. Zucchini, 1997, *Hidden Markov and Other Models for Discrete-Valued Time Series*, Taylor & Francis.