10

Finite Markov Mixture Modeling

10.1 Introduction

In this and the following chapters, finite mixture models are extended to deal with time series data that exhibit dependence over time. Broadly speaking, this is achieved by substituting the discrete latent indicator S_i introduced as an allocation variable for finite mixture models by a hidden Markov chain. This leads to a surprisingly rich class of nonlinear time series models that solve a variety of interesting problems in applied time series analysis, as demonstrated in Chapter 12.

Section 10.2 starts with the definition of a finite Markov mixture distribution, whose properties are studied in some detail. Section 10.3 introduces the basic Markov switching model and deals with its extensions. The problem of econometric estimation of a Markov switching model from an observed time series is then discussed in Chapter 11.

10.2 Finite Markov Mixture Distributions

Let $\{y_t, t = 1, \ldots, T\}$ denote a time series of T univariate observations taking values in a sampling space $\mathcal{Y}$ which may be either discrete or continuous. As common in time series analysis, $\{y_t, t = 1, \ldots, T\}$ is considered to be the realization of a stochastic process $\{Y_t\}_{t=1}^{T}$. Modeling is based on special cases from the class of doubly stochastic time series models (Tjøstheim, 1986) that have been found to be very useful for applied time series analysis.

It is assumed that the probability distribution of the stochastic process Y_t depends on the realizations of a hidden discrete stochastic process S_t. The stochastic process Y_t is directly observable, whereas S_t is a latent random process that is observable only indirectly through the effect it has on the realizations of Y_t. A simple example is the hidden Markov chain model $Y_t = \mu_{S_t} + \varepsilon_t$, where ε_t is a zero-mean white noise process with variance σ^2.

10.2.1 Basic Definitions

We start with specifying the properties of the hidden process $\{S_t\}_{t=0}^T$, which is assumed to be a discrete-time process with finite state space $\{1, \ldots, K\}$ that obeys the following condition **S4**.

S4 S_t is an irreducible, aperiodic Markov chain starting from its ergodic distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$:

$$\Pr(S_0 = k|\boldsymbol{\xi}) = \eta_k.$$

The stochastic properties of S_t are sufficiently described by the $(K \times K)$ transition matrix $\boldsymbol{\xi}$, where each element ξ_{jk} of $\boldsymbol{\xi}$ is equal to the transition probability from state j to state k:

$$\xi_{jk} = \Pr(S_t = k|S_{t-1} = j), \qquad \forall j, k \in \{1, \ldots, K\}.$$

Evidently, the jth row of the transition matrix $\boldsymbol{\xi}$ defines, for all $t = 1, \ldots, T$, the conditional distribution of S_t given the information that S_{t-1} is in state j. We sometimes use the notation $\boldsymbol{\xi}_{j\cdot}$ to refer to row j. All elements of $\boldsymbol{\xi}$ are nonnegative and the elements of each row sum to 1:

$$\begin{aligned} &\xi_{jk} \geq 0, \qquad \forall j, k \in \{1, \ldots, K\}, \\ &\sum_{k=1}^{K} \xi_{jk} = 1, \qquad \forall j = 1, \ldots, K. \end{aligned} \tag{10.1}$$

$\boldsymbol{\xi} = (\boldsymbol{\xi}_{1\cdot}, \ldots, \boldsymbol{\xi}_{K\cdot})$ takes values in the product space $(\mathcal{E}_K)^K$, where $\mathcal{E}_K$ is the unit simplex defined in Subsection 1.2.1. Further assumptions about $\boldsymbol{\xi}$ are necessary to fulfill condition **S4**; see Subsection 10.2.2 for more details.

We continue with describing how the distribution of Y_t depends on S_t. Let $\mathcal{T}(\boldsymbol{\theta})$ be a parametric distribution family, defined over a sampling space $\mathcal{Y}$ which may be either discrete or continuous, with density $p(y|\boldsymbol{\theta})$, indexed by a parameter $\boldsymbol{\theta} \in \Theta$. Let $\{Y_t\}_{t=1}^T$ be a sequence of random variables that depend on $\{S_t\}_{t=0}^T$ in the following way.

Y4 Conditional on knowing $\mathbf{S} = (S_0, \ldots, S_T)$, the random variables $Y_1, \ldots, Y_T$ are stochastically independent. For each $t \geq 1$, the distribution of Y_t arises from one out of K distributions $\mathcal{T}(\boldsymbol{\theta}_1), \ldots, \mathcal{T}(\boldsymbol{\theta}_K)$, depending on the state of S_t:

$$Y_t|S_t = k \sim \mathcal{T}(\boldsymbol{\theta}_k).$$

Hidden indicators comparable to S_t have been introduced also for a finite mixture model, using the symbol S_i. The original definition of a mixture distribution in Section 1.2, however, started with the marginal distribution of Y_i without introducing the latent indicator S_i right from the beginning.

For the doubly stochastic process $\{S_t, Y_t\}_{t=1}^T$ obeying conditions **S4** and **Y4** it is rather easy to derive the marginal distribution of Y_t:

$$p(y_t|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(y_t|S_t = k, \boldsymbol{\vartheta}) \Pr(S_t = k|\boldsymbol{\vartheta}).$$

Because S_t is a stationary Markov chain and the conditional distribution of Y_t given $S_t = k$ has density $p(y_t|\boldsymbol{\theta}_k)$, one obtains that the unconditional distribution of Y_t is a finite mixture of $\mathcal{T}(\boldsymbol{\theta})$ distribution with the ergodic probabilities $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ acting as weight distribution (Baum et al., 1970):

$$p(y_t|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(y_t|\boldsymbol{\theta}_k)\eta_k. \tag{10.2}$$

Hence the process Y_t is said to be generated by a finite Markov mixture of $\mathcal{T}(\boldsymbol{\theta})$ distributions. Stationarity of Y_t is evident from (10.2). Furthermore such a process is autocorrelated (see Subsection 10.2.4), which is an important difference to a (standard) finite mixture of $\mathcal{T}(\boldsymbol{\theta})$ distributions, which produces sequences of independent random variables.

One early example of a finite Markov mixture distribution is the hidden Markov chain model (Baum and Petrie, 1966), where Y_t is a discrete random signal taking one out of D values $\{1, \ldots, D\}$ according to a discrete probability distribution, which depends on the state of S_t:

$$\Pr(Y_t = l|S_t = k) = \pi_{k,l},$$

for $k = 1, \ldots, K$ and $l = 1, \ldots, D$. The transition matrix $\boldsymbol{\xi}$ as well as the matrix $\boldsymbol{\Pi} = (\pi_{k,l})$ is assumed to be unknown and has to be recovered from observations $\mathbf{y} = (y_1, \ldots, y_T)$ of the process $\{Y_t\}_{t=1}^T$, whereas S_t is unobserved.

Another early example is a Markov mixture of normal distributions (Baum et al., 1970), where Y_t is a discrete signal observed with noise:

$$Y_t = \begin{cases} \mu_1 + \varepsilon_t, \ \varepsilon_t \sim \mathcal{N}\left(0, \sigma_1^2\right), & S_t = 1, \\ \vdots \\ \mu_K + \varepsilon_t, \ \varepsilon_t \sim \mathcal{N}\left(0, \sigma_K^2\right), & S_t = K. \end{cases}$$

Many more examples appear throughout the remaining chapters.

The mathematical properties of a process generated by a finite Markov mixture distribution have been studied for specific processes obeying conditions **Y4** and **S4** such as hidden Markov chain models (Blackwell and Koopmans, 1957; Heller, 1965), white noise driven by a hidden Markov chain (Francq and Roussignol, 1997), discrete-valued time series generated by a hidden Markov chain (MacDonald and Zucchini, 1997), Markov mixtures of normal distributions (Krolzig, 1997), and Markov mixtures of more general

location-scale families, where $Y_t = \mu_{S_t} + \sigma_{S_t}\varepsilon_t$, with ε_t being an i.i.d. process (Timmermann, 2000).

After a short introduction into irreducible, aperiodic Markov chains in Subsection 10.2.2 these results are summarized for arbitrary processes Y_t being generated by a Markov mixture obeying conditions **Y4** and **S4**.

10.2.2 Irreducible Aperiodic Markov Chains

In this subsection we briefly review properties of irreducible aperiodic Markov chains, focusing on results that are needed later on. For a more detailed survey we refer to Karlin and Taylor (1975).

Let S_t be a homogeneous first-order Markov chain with transition matrix $\boldsymbol{\xi}$, where the elements of $\boldsymbol{\xi}$ are unconstrained apart from the natural constraints defined in (10.1). The transition matrix $\boldsymbol{\xi}$ plays a prominent role in understanding the properties of the corresponding Markov chain.

Any probability distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ that fulfills the invariance property

$$\boldsymbol{\xi}^{'}\boldsymbol{\eta} = \boldsymbol{\eta}, \tag{10.3}$$

is called an invariant distribution of S_t. The practical importance of the invariant distribution for the Markov chain S_t is the following. Assume that at time $t-1$ the states of S_{t-1} are drawn from an invariant distribution of $\boldsymbol{\xi}$. Then the following holds $\forall k = 1, \ldots, K$,

$$\begin{aligned}\Pr(S_t = k|\boldsymbol{\xi}) &= \sum_{j=1}^{K} \Pr(S_t = k|S_{t-1} = j, \boldsymbol{\xi})\Pr(S_{t-1} = j|\boldsymbol{\xi}) \\ &= \sum_{j=1}^{K} \xi_{jk}\eta_j = \eta_k.\end{aligned}$$

Therefore the states of S_t are again drawn from $\boldsymbol{\eta}$, and so on for $S_{t+1}, \ldots$.

It is possible to show that such an invariant distribution exists for any finite Markov chain. By rewriting the constraint (10.1) as

$$\boldsymbol{\xi}_{j\cdot}\mathbf{1}_{K\times 1} = 1, \qquad \forall j = 1, \ldots, K,$$

where $\boldsymbol{\xi}_{j\cdot}$ refers to row j of $\boldsymbol{\xi}$, and $\mathbf{1}_{K\times 1}$ is a column vector of ones, it becomes apparent, that for any transition matrix $\boldsymbol{\xi}$ one of the eigenvalues is equal to 1:

$$\boldsymbol{\xi}\mathbf{1}_{K\times 1} = 1 \times \mathbf{1}_{K\times 1}.$$

By rewriting (10.3) as

$$\boldsymbol{\eta}^{'}\boldsymbol{\xi} = \boldsymbol{\eta}^{'} \times 1,$$

it becomes apparent that, formally, $\boldsymbol{\eta}$ is the (suitably normalized) left-hand eigenvector of $\boldsymbol{\xi}$, associated with the eigenvalue 1.

The invariant distribution, however, is not unique for arbitrary transition matrices $\boldsymbol{\xi} \in (\mathcal{E}_K)^K$; consider, for instance, the transition matrix $\boldsymbol{\xi} = \mathbf{I}_K$, for which any arbitrary probability distribution will be invariant. An outstanding subset in the class $(\mathcal{E}_K)^K$ contains transition matrices for which this invariant distribution is unique and, additionally, the distribution of S_t converges to this invariant distribution, regardless of the state of S_0. Such a Markov chain is called an ergodic Markov chain, and the invariant distribution $\boldsymbol{\eta}$ is called the ergodic distribution of the Markov chain.

Necessary restrictions on $\boldsymbol{\xi}$ to achieve ergodicity may be defined in terms of properties of $\boldsymbol{\xi}^h = \boldsymbol{\xi} \cdots \boldsymbol{\xi}$, the hth power of the transition matrix $\boldsymbol{\xi}$. $\boldsymbol{\xi}^h$ determines the long-run behavior of the Markov chain in terms of the h-step ahead predictive distribution $\Pr(S_{t+h} = l | S_t = k, \boldsymbol{\xi})$ of S_{t+h} given $S_t = k$:

$$\Pr(S_{t+h} = l | S_t = k, \boldsymbol{\xi}) = (\boldsymbol{\xi}^h)_{kl}, \tag{10.4}$$

where $(\boldsymbol{\xi}^h)_{kl}$ is the element (k, l) of $\boldsymbol{\xi}^h$. (10.4) is obvious for $h = 1$ from the definition of $\boldsymbol{\xi}$. For $h > 1$, (10.4) is easily derived by induction:

$$\begin{aligned} &\Pr(S_{t+h} = l | S_t = k, \boldsymbol{\xi}) \\ &= \sum_{j=1}^{K} \Pr(S_{t+h} = l | S_{t+h-1} = j, \boldsymbol{\xi}) \Pr(S_{t+h-1} = j | S_t = k, \boldsymbol{\xi}) \\ &= \sum_{j=1}^{K} \xi_{jl} (\boldsymbol{\xi}^{h-1})_{kj} = (\boldsymbol{\xi}^h)_{kl}. \end{aligned}$$

Uniqueness of the invariant distribution follows for any transition matrix that leads to an irreducible Markov chain. Irreducibility means that starting S_t from an arbitrary state $k \in \{1, \ldots, K\}$ any state $l \in \{1, \ldots, K\}$ must be reachable in finite time, or in terms of $(\boldsymbol{\xi}^h)_{kl}$:

$$\forall (k, l) \in \{1, \ldots, K\} \quad \Rightarrow \quad \exists h(k, l) : (\boldsymbol{\xi}^{h(k,l)})_{kl} > 0. \tag{10.5}$$

It follows that any transition matrix $\boldsymbol{\xi}$ where all elements ξ_{jk} are positive leads to irreducibility and uniqueness of the invariant distribution. More generally, irreducibility follows if $(\boldsymbol{\xi}^h)_{kl} > 0$ for some $h \geq 1$, independent of k, l. If any element $(\boldsymbol{\xi}^h)_{kl}$ is 0 for all $h \geq 1$, then the Markov chain is reducible; consider, for instance, the following transition matrix

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & 1 - \xi_{11} \\ 0 & 1 \end{pmatrix}, \tag{10.6}$$

which reappears in Subsection 10.3.3 in the context of change-point modeling. It is easily verified that this transition matrix leads to a reducible Markov chain, in as much as for all $h \geq 1$:

$$\boldsymbol{\xi}^h = \begin{pmatrix} \xi_{11}^h & 1-\xi_{11}^h \\ 0 & 1 \end{pmatrix}.$$

Solving (10.3) for $K = 2$ leads to the following invariant probabilities,

$$\begin{aligned} \eta_1 &= \frac{1-\xi_{22}}{(1-\xi_{11})+(1-\xi_{22})} = \frac{\xi_{21}}{\xi_{12}+\xi_{21}}, \\ \eta_2 &= \frac{1-\xi_{11}}{(1-\xi_{11})+(1-\xi_{22})} = \frac{\xi_{12}}{\xi_{12}+\xi_{21}}. \end{aligned} \tag{10.7}$$

For a Markov chain with $\xi_{11} = \xi_{22}$, the invariant probability distribution is uniform: $\eta_1 = \eta_2 = 0.5$; $\xi_{11} > \xi_{22}$ favors state 1: $\eta_1 > \eta_2$, whereas $\xi_{11} < \xi_{22}$ favors state 2: $\eta_1 < \eta_2$.

For $K > 2$, some numerical method has to be used for solving (10.3). A closed-form expression for the invariant probability distribution $\boldsymbol{\eta}$ in terms of the transition matrix $\boldsymbol{\xi}$ is derived in Hamilton (1994b, Section 22.2). Define a matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_K - \boldsymbol{\xi}' \\ \mathbf{1}_{1\times K} \end{pmatrix}, \tag{10.8}$$

with $\mathbf{I}_K$ being the identity matrix with K rows and $\mathbf{1}_{1\times K}$ being a row vector of ones. Then $\boldsymbol{\eta}$ is given as the $(K+1)$th column of the matrix $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$:

$$\boldsymbol{\eta} = \left((\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\right)_{\cdot,K+1}. \tag{10.9}$$

Now let us turn to the distribution $\Pr(S_t|\boldsymbol{\xi})$ of a Markov chain S_t, starting with S_0 being drawn from a certain probability distribution. If the states of S_0 are drawn from the invariant distribution $\boldsymbol{\eta}$ of $\boldsymbol{\xi}$, then by the invariance property $\Pr(S_t|\boldsymbol{\xi})$ is equal $\boldsymbol{\eta}$ for all $t \geq 1$, but what happens if S_0 is drawn from a different distribution or is assumed to be a fixed starting value? Consider, for instance, the following irreducible transition matrix

$$\boldsymbol{\xi} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{10.10}$$

This matrix is an example of a doubly stochastic matrix where both the row and the column sums are equal to 1:

$$\sum_{k=1}^{K} \xi_{jk} = 1, \qquad \sum_{j=1}^{K} \xi_{jk} = 1.$$

For such matrices the uniform distribution, $\eta_k = 1/K$, is an invariant distribution:

$$\sum_{j=1}^{K} \xi_{jk}\eta_j = \frac{1}{K}\sum_{j=1}^{K} \xi_{jk} = \frac{1}{K}, \qquad \forall k = 1, \ldots, K.$$

Because (10.10) is irreducible, the uniform distribution is the unique invariant distribution; the distribution $\Pr(S_t|\boldsymbol{\xi})$, however, does not converge to the invariant distribution if S_0 is started with a different distribution, such as the degenerate distribution $\Pr(S_0 = 1) = 1$, because

$$\begin{aligned}
&\Pr(S_t = 1|S_0 = 1, \boldsymbol{\xi}) = 1, \qquad &&\text{iff } t = 3m + 1, m \in \{1, 2, 3, \ldots\},\\
&\Pr(S_t = 2|S_0 = 1, \boldsymbol{\xi}) = 1, \qquad &&\text{iff } t = 3m + 2, m \in \{1, 2, 3, \ldots\},\\
&\Pr(S_t = 3|S_0 = 1, \boldsymbol{\xi}) = 1, \qquad &&\text{iff } t = 3m, m \in \{1, 2, 3, \ldots\}.
\end{aligned}$$

The main reason for this failure of convergence is that the transition matrix (10.10) is periodic and captures a kind of seasonal pattern.

Ergodicity of a Markov chain with transition matrix $\boldsymbol{\xi}$ holds, if the Markov chain is aperiodic. Aperiodicity is defined as the absence of periodicity such as the one observed in the transition matrix (10.10). Consider, for each state k, all periods n for which the transition probability $\Pr(S_{t+n} = k|S_t = k, \boldsymbol{\xi}) = (\boldsymbol{\xi}^n)_{kk}$ is positive. The period of a state is the greatest common divisor (GCD) of all periods n. A Markov chain is aperiodic, if the period of each state is equal to one:

$$\text{GCD}\{n \geq 1 : (\boldsymbol{\xi}^n)_{kk} > 0\} = 1, \qquad \forall k \in \{1, \ldots, K\}.$$

A Markov chain is aperiodic if all diagonal elements of $\boldsymbol{\xi}$ are positive.

Ergodicity of a Markov chain implies that the distribution $\Pr(S_t|\boldsymbol{\xi}, S_0 = k)$ which is equal to the kth row $(\boldsymbol{\xi}^h)_{k\cdot}$ of $\boldsymbol{\xi}^h$ converges to the ergodic distribution, regardless of the state k of S_0:

$$\lim_{h\to\infty} (\boldsymbol{\xi}^h)_{k\cdot} = \boldsymbol{\eta}'.$$

For understanding Markov mixture models it is helpful to know if this convergence is fast or if the Markov chain S_t is persistent, meaning that the state of S_t is mainly defined by the state of S_{t-1}. It turns out that the second largest eigenvalues of $\boldsymbol{\xi}$ play a crucial role in this respect.

Consider, for instance, a two-state Markov chain, where

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & 1 - \xi_{11} \\ 1 - \xi_{22} & \xi_{22} \end{pmatrix}.$$

A two-state Markov chain is ergodic if $0 < \xi_{11} + \xi_{22} < 2$. The eigenvalues are obtained from

$$\begin{vmatrix} \xi_{11} - \lambda & 1 - \xi_{11} \\ 1 - \xi_{22} & \xi_{22} - \lambda \end{vmatrix} = (\lambda - 1)(\lambda - (\xi_{11} + \xi_{22} - 1)) = 0.$$

Apart from $\lambda = 1$, the other eigenvalue is equal to:

$$\lambda = \xi_{11} + \xi_{22} - 1 = \xi_{11} - \xi_{21}. \tag{10.11}$$

For $K = 2$ a simple representation of $\boldsymbol{\xi}^h$ in terms of the ergodic probability distribution is possible (Hamilton, 1994a, p.683):

$$\boldsymbol{\xi}^h = \begin{pmatrix} \eta_1 & \eta_2 \\ \eta_1 & \eta_2 \end{pmatrix} + \lambda^h \begin{pmatrix} \eta_2 & -\eta_2 \\ -\eta_1 & \eta_1 \end{pmatrix}, \tag{10.12}$$

with λ being the second eigenvalue derived in (10.11) which demonstrates that persistence of S_t is higher, the closer λ is to 1.

Persistence is also related to the issue of duration of a certain state. Given that the Markov chain S_t is currently in state j, the duration D_j of that state is a random variable following a geometric distribution with parameter $1-\xi_{jj}$ (see Appendix A.1.7),

$$\begin{aligned} &\Pr(D_j = l | S_t = j) = \Pr(S_{t+1} = j, \ldots, S_{t+l-1} = j, S_{t+l} \neq j | S_t = j) \\ &= \prod_{m=1}^{l-1} \Pr(S_{t+m} = j | S_{t+m-1} = j) \Pr(S_{t+l} \neq j | S_{t+l-1} = j) \\ &= \xi_{jj}^{l-1}(1 - \xi_{jj}). \end{aligned}$$

Therefore the expected duration of state j is given by

$$\mathrm{E}(D_j) = \frac{1}{1 - \xi_{jj}}. \tag{10.13}$$

Two interesting conclusions may be drawn from (10.13). First, the expected duration of state j is longer the closer the persistence probability ξ_{jj} is to 1. Second, if the persistence probabilities differ in the various states, then also the expected duration of the state differs across states. Therefore Markov mixture distributions are able to capture asymmetry over time as observed for economic time series such as unemployment (Neftçi, 1984) and GDP, investment, and industrial production (Falk, 1986) over the business cycle.

10.2.3 Moments of a Markov Mixture Distribution

Because the unconditional distribution of a random process Y_t, being generated by a Markov mixture of $\mathcal{T}(\boldsymbol{\theta})$-distribution is a standard finite mixture of $\mathcal{T}(\boldsymbol{\theta})$-distribution with the ergodic probabilities acting as weights, the expectation of any function $h(Y_t)$ of Y_t is given by the results of Subsection 1.2.4, where $\boldsymbol{\eta}$ is substituted by the ergodic distribution of S_t.

From Subsection 1.2.4 it is known that standard finite mixture distributions are able to generate probability distributions with asymmetry and fat tails. Timmermann (2000) studied finite Markov mixture distributions taken from a location-scale family and demonstrated that the introduction of Markovian dependence into the hidden indicator S_t even increases the scope for asymmetry and fat tails in the generated process.

Moments of a Markov Mixture of Two Normal Distributions

More explicit results are given for a Markov mixture of two normal distributions:

$$Y_t = \begin{cases} \mu_1 + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}\left(0, \sigma_1^2\right), & S_t = 1, \\ \mu_2 + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}\left(0, \sigma_2^2\right), & S_t = 2. \end{cases}$$

The unconditional distribution of Y_t is given by a mixture of two normal distributions:

$$p(y_t|\boldsymbol{\vartheta}) = \eta_1 f_N(y_t; \mu_1, \sigma_1^2) + \eta_2 f_N(y_t; \mu_2, \sigma_2^2), \tag{10.14}$$

where the ergodic probabilities η_1 and η_2 are given by (10.7). The marginal distribution (10.14) exhibits nonnormality as long as either $\mu_1 \neq \mu_2$ or $\sigma_1^2 \neq \sigma_2^2$. Multimodality of the marginal distribution is possible for appropriate choices of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \xi_{11}, \xi_{21})$ and could be checked for a given parameter using the results of Subsection 1.2.2.

From Subsection 1.2.4 the following coefficient of skewness results,

$$\frac{\mathrm{E}((Y_t - \mu)^3|\boldsymbol{\vartheta})}{\mathrm{E}((Y_t - \mu)^2|\boldsymbol{\vartheta})^{3/2}} = \eta_1\eta_2(\mu_1 - \mu_2)\frac{3(\sigma_2^2 - \sigma_1^2)^2 + (\eta_2 - \eta_1)(\mu_2 - \mu_1)^2}{\sigma^3},$$

with $\mu = \mathrm{E}(Y_t|\boldsymbol{\vartheta})$ and $\sigma^2 = \mathrm{Var}(Y_t|\boldsymbol{\vartheta})$ being the mean and variance of the mixture distribution (10.14):

$$\begin{aligned} \mu &= \eta_1\mu_1 + \eta_2\mu_2, \\ \sigma^2 &= \eta_1\sigma_1^2 + \eta_2\sigma_2^2 + \eta_1\eta_2(\mu_2 - \mu_1)^2. \end{aligned}$$

Skewness in the marginal distribution will be present whenever both the means and the variances are different. For a model where the means are the same, no skewness is present. If the variances are the same and the means are different, skewness is possible only iff $\eta_1 \neq \eta_2$. Thus, for a Markov mixture model with different means but equal variances, asymmetry is introduced into the marginal distribution only through asymmetry in the persistence probabilities, namely $\xi_{11} \neq \xi_{22}$.

From Subsection 1.2.4, excess kurtosis is given by

$$\frac{\mathrm{E}((Y_t - \mu)^4|\boldsymbol{\vartheta})}{\mathrm{E}((Y_t - \mu)^2|\boldsymbol{\vartheta})^2} - 3 = \eta_1\eta_2\frac{3(\sigma_2^2 - \sigma_1^2)^2 + c(\mu_1, \mu_2)}{\sigma^4}, \tag{10.15}$$

where $c(\mu_1, \mu_2) = 6(\eta_1 - \eta_2)(\sigma_2^2 - \sigma_1^2)(\mu_2 - \mu_1)^2 + (\mu_2 - \mu_1)^4(1 - 6\eta_1\eta_2)$; see also Timmermann (2000, Corollary 1). Therefore if $\mu_1 = \mu_2$, the marginal distribution has fatter tails than a normal distribution as long as $\sigma_1^2 \neq \sigma_2^2$.

10.2.4 The Autocorrelation Function of a Process Generated by a Markov Mixture Distribution

A finite Markov mixture distribution generates an autocorrelated process Y_t where the autocorrelation strongly depends on the persistence of S_t. The autocorrelation function of Y_t is defined in the usual way as

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \frac{\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta}) - \mu^2}{\sigma^2}, \tag{10.16}$$

with $\mu = \mathrm{E}(Y_t|\boldsymbol{\vartheta})$ and $\sigma^2 = \mathrm{Var}(Y_t|\boldsymbol{\vartheta})$ being the unconditional moments and

$$\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta}) = \int y_t y_{t+h} p(y_t, y_{t+h}|\boldsymbol{\vartheta}) dy_t dy_{t+h}.$$

MacDonald and Zucchini (1997) derive the autocorrelation function for hidden Markov chain models for time series of counts, whereas Krolzig (1997), Rydén et al. (1998), and Timmermann (2000) consider continuous data. In the following we provide results for arbitrary processes obeying conditions **S4** and **Y4**.

To this aim it is useful to give an explicit form for the density $p(y_t, y_{t+h}|\boldsymbol{\vartheta})$ of the joint unconditional distribution of Y_t and Y_{t+h}:

$$\begin{aligned} p(y_t, y_{t+h}|\boldsymbol{\vartheta}) = \sum_{k,l=1}^{K} & p(y_t|S_t = k, \boldsymbol{\vartheta}) p(y_{t+h}|S_{t+h} = l, \boldsymbol{\vartheta}) \\ & \times \Pr(S_{t+h} = l|S_t = k, \boldsymbol{\xi}) \Pr(S_t = k|\boldsymbol{\xi}). \end{aligned} \tag{10.17}$$

The predictive distribution $\Pr(S_{t+h} = l|S_t = k, \boldsymbol{\xi})$ is given by (10.4), and (10.17) reduces to

$$p(y_t, y_{t+h}|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(y_t|\boldsymbol{\theta}_k)\eta_k \sum_{l=1}^{K} p(y_{t+h}|\boldsymbol{\theta}_l)(\boldsymbol{\xi}^h)_{kl}, \tag{10.18}$$

where $(\boldsymbol{\xi}^h)_{kl}$ is the element (k, l) of the hth power of the transition matrix $\boldsymbol{\xi}$. Therefore $\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta})$ is given by

$$\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \mu_k \sum_{l=1}^{K} (\boldsymbol{\xi}^h)_{kl} \mu_l, \tag{10.19}$$

and the autocorrelation function results from (10.16):

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \frac{\displaystyle\sum_{k=1}^{K} \mu_k \eta_k \sum_{l=1}^{K} \mu_l (\boldsymbol{\xi}^h)_{kl} - \mu^2}{\sigma^2}.$$

Because the process Y_t is uncorrelated conditional on knowing S_t, the autocorrelation function depends on h only through $\boldsymbol{\xi}^h$, and autocorrelation in the marginal process Y_t, where S_t is unknown, enters through persistence in S_t, only. Note that Y_t, in contrast to S_t, is no longer a Markov process of first order.

Autocorrelation for a Two-State Model

From the specific form of $\boldsymbol{\xi}^h$ given in (10.12), the following autocorrelation function results for any two-state finite Markov mixture model,

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \frac{\eta_1\eta_2(\mu_1-\mu_2)^2}{\sigma^2}\lambda^h, \tag{10.20}$$

with $\lambda = \xi_{11} - \xi_{21}$ being the second eigenvalue of $\boldsymbol{\xi}$.

No autocorrelation in Y_t is present if $\mu_1 = \mu_2$. Otherwise, autocorrelation of Y_t is caused through the hidden Markov chain S_t, whenever $\xi_{11} \neq \xi_{21}$. The process Y_t exhibits positive autocorrelation provided that $\xi_{11} > \xi_{21}$, otherwise negative autocorrelation results. An equivalent criterion is to check if $\xi_{11}+\xi_{22}$ is larger or smaller than 1.

Relation to ARMA Models

There exists a close relationship between Markov mixture models and nonnormal ARMA models. For a two-state Markov mixture model, for instance, the autocorrelation function of Y_t given in (10.20) fulfills, for $h > 1$, the following recursion,

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \lambda\rho_{Y_t}(h-1|\boldsymbol{\vartheta}),$$

and corresponds to the autocorrelation function of an ARMA$(1,1)$ process, whereas the nonnormality of the unconditional distribution of Y_t is preserved through the mixture distribution. In general, Poskitt and Chung (1996) proved for a univariate K-state hidden Markov chain $Y_t = \mu_{S_t} + u_t$ the existence of an ARMA$(K-1, K-1)$ representation with a homogeneous zero-mean white noise process.

10.2.5 The Autocorrelation Function of the Squared Process

An interesting feature of any finite Markov mixture model is that it generates processes Y_t, with Y_t^2 being autocorrelated. This is of particular interest when Markov mixture models are applied to financial time series; see Section 12.5. Timmermann (2000, Proposition 5) derived the autocorrelation function for a Markov mixture based on the continuous location-scale family. It is quite easy to generalize these results to any process obeying conditions **Y4** and **S4**.

The autocorrelation function of Y_t^2 is defined as

$$\rho_{Y_t^2}(h|\boldsymbol{\vartheta}) = \frac{\mathrm{E}(Y_t^2Y_{t+h}^2|\boldsymbol{\vartheta}) - \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2}{\mathrm{E}(Y_t^4|\boldsymbol{\vartheta}) - \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2}, \tag{10.21}$$

where $\mathrm{E}(Y_t^2|\boldsymbol{\vartheta}) = \sum_{k=1}^K \mathrm{E}(Y_t^2|\boldsymbol{\theta}_k)\eta_k$, and $\mathrm{E}(Y_t^4|\boldsymbol{\vartheta}) = \sum_{k=1}^K \mathrm{E}(Y_t^4|\boldsymbol{\theta}_k)\eta_k$, and $\mathrm{E}(Y_t^2Y_{t+h}^2|\boldsymbol{\vartheta})$ is obtained from (10.18) as

$$\mathrm{E}(Y_t^2 Y_{t+h}^2|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \mathrm{E}(Y_t^2|\boldsymbol{\theta}_k) \sum_{l=1}^{K} \mathrm{E}(Y_{t+h}^2|\boldsymbol{\theta}_l)(\boldsymbol{\xi}^h)_{kl}, \qquad (10.22)$$

with $(\boldsymbol{\xi}^h)_{kl}$ being the element (k,l) of the hth power of the transition matrix $\boldsymbol{\xi}$. Although the process Y_t^2 is uncorrelated conditional on knowing S_t, autocorrelation in Y_t^2 enters through persistence in S_t.

Autocorrelation in the Squared Process for a Two-State Model

We provide here further details for a Markov mixture of two normal distributions. From the general autocorrelation function of Y_t^2 given by (10.21), together with the representation of the transition matrix $\boldsymbol{\xi}$ of a two-state Markov model as in (10.12), one obtains:

$$\mathrm{E}(Y_t^2 Y_{t+h}^2|\boldsymbol{\vartheta}) = \mathrm{E}(Y_t^2|\boldsymbol{\vartheta}) + \eta_1\eta_2(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2)^2\lambda^h, \qquad (10.23)$$

with $\lambda = \xi_{11} - \xi_{21}$ being the second eigenvalue of $\boldsymbol{\xi}$. Therefore:

$$\rho_{Y_t^2}(h|\boldsymbol{\vartheta}) = \frac{\eta_1\eta_2(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2)^2}{\mathrm{E}(Y_t^4|\boldsymbol{\vartheta}) - \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2}\lambda^h. \qquad (10.24)$$

The squared process exhibits positive autocorrelation provided that $\xi_{11} > \xi_{21}$, otherwise if $\xi_{11} < \xi_{21}$ negative autocorrelation will result. An equivalent criterion is to check if $\xi_{11} + \xi_{22}$ is larger or smaller than 1. Interestingly, state dependent variances are neither necessary nor sufficient for autocorrelation in the squared process. Even if $\sigma_1^2 = \sigma_2^2$, the marginal process shows conditional heteroscedasticity, as long as S_t does not degenerate to an i.i.d. process. On the other hand, if $\xi_{11} = \xi_{21}$, no autocorrelation in the squared returns is present, even if $\sigma_1^2 \neq \sigma_2^2$.

By comparing the autocorrelation of Y_t^2, given by (10.24), with the autocorrelation of Y_t, given by (10.20), we find that a Markov mixture of two normal distributions with $\mu_1 = \mu_2$ will produce an uncorrelated process without skewness in the marginal distribution, whereas Y_t^2 is correlated and the marginal distribution has fat tails, as long as $\sigma_1^2 \neq \sigma_2^2$. As for other models that capture autocorrelation in the squared process, such as the GARCH model (Bollerslev, 1986), differences in the variances alone are insufficient to capture asymmetry in the marginal distribution.

10.2.6 The Standard Finite Mixture Distribution as a Limiting Case

Any standard finite mixture of $\mathcal{T}(\boldsymbol{\theta})$-distributions defined in Chapter 1 may be thought of as that limiting case of a finite Markov mixture of $\mathcal{T}(\boldsymbol{\theta})$-distribution where S_t is an i.i.d. random sequence, in which case the transition probabilities from state j to state k are equal to $\Pr(S_t = k|S_{t-1} = j) = \Pr(S_t = k) = \eta_k$.

Thus a random variable Y_t drawn from a standard finite mixture of $\mathcal{T}(\boldsymbol{\theta})$-distribution with weight distribution $\boldsymbol{\eta}$ is observationally equivalent with a process Y_t generated by a finite Markov mixture of $\mathcal{T}(\boldsymbol{\theta})$-distributions where all rows of the transition matrix of S_t are identical to $\boldsymbol{\eta}$:

$$\boldsymbol{\xi} = \begin{pmatrix} \eta_1 & \cdots & \eta_K \\ \vdots & & \vdots \\ \eta_1 & \cdots & \eta_K \end{pmatrix}.$$

In this case the transition matrix $\boldsymbol{\xi}$ is idempotent, $\boldsymbol{\xi}^h = \boldsymbol{\xi}$ for all $h \geq 1$, and (10.19) reduces to

$$\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \mu_k \sum_{l=1}^{K} \xi_{kl} \mu_l = \mu^2.$$

Thus the autocorrelation $\rho_{Y_t}(h|\boldsymbol{\vartheta})$ of Y_t, given by (10.16), is equal to 0 for $h > 1$. Similarly, (10.22) reduces to

$$\mathrm{E}(Y_t^2 Y_{t+h}^2|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \mathrm{E}(Y_t^2|\boldsymbol{\theta}_k) \sum_{l=1}^{K} \mathrm{E}(Y_{t+h}^2|\boldsymbol{\theta}_l) \xi_{kl} = \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2,$$

and the autocorrelation $\rho_{Y_t^2}(h|\boldsymbol{\vartheta})$ of Y_t^2, given by (10.21), is equal to 0 for $h > 1$.

10.2.7 Identifiability of a Finite Markov Mixture Distribution

For a finite Markov mixture distribution one has to distinguish between the same three types of nonidentifiability that have been discussed for a standard finite mixture distribution in Section 1.3. There exists nonidentifiability due to invariance to relabeling the states of the hidden Markov chain as well as generic nonidentifiability.

Consider all $s = 1, \ldots, K!$ different permutations $\rho_s : \{1, \ldots, K\} \to \{1, \ldots, K\}$, where the value $\rho_s(k)$ is assigned to each value $k \in \{1, \ldots, K\}$. Let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\xi})$ be an arbitrary point in the parameter space $\Theta_K = \Theta^K \times (\mathcal{E}_K)^K$, and define a subset $\mathcal{U}^P(\boldsymbol{\vartheta}) \subset \Theta_K$ by

$$\mathcal{U}^P(\boldsymbol{\vartheta}) = \bigcup_{s=1}^{K!} \{\boldsymbol{\vartheta}^\star \in \Theta_K : \boldsymbol{\vartheta}^\star = (\boldsymbol{\theta}_{\rho_s(1)}, \ldots, \boldsymbol{\theta}_{\rho_s(K)}, \boldsymbol{\xi}^{\rho_s})\}, \tag{10.25}$$

where $\boldsymbol{\xi}^{\rho_s}$ is related to $\boldsymbol{\xi}$ by permuting the rows and the column in the same fashion:

$$\xi_{jk}^{\rho_s} = \xi_{\rho_s(j),\rho_s(k)}, \qquad \forall j, k \in \{1, \ldots, K\}. \tag{10.26}$$

Then evidently, all points in $\mathcal{U}^P(\boldsymbol{\vartheta})$ generate the same Markov mixture distribution, however, with a different labeling of the states of the hidden Markov chain.

A weak inequality constraint, similar to the one discussed for finite mixtures in Subsection 1.3.3 requiring that the state-specific parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_l$ differ in *at least one* element, which need not be the same for all states, will rule out these identifiability problems.

Blackwell and Koopmans (1957) is an early reference addressing generic identifiability problems for some special hidden Markov chain models, where Y_t is a discrete signal. Petrie (1969) proved generic identifiability for hidden Markov chain models, where the observed process Y_t takes values in a finite set. Identifiability for rather general finite Markov mixtures is addressed in Leroux (1992b).

One necessary condition for generic identifiability of a Markov mixture of $\mathcal{T}(\boldsymbol{\theta})$-distributions is that a standard finite mixture of $\mathcal{T}(\boldsymbol{\theta})$-distributions is generically identifiable; see again Subsection 1.3.4. A second necessary condition is that the hidden Markov chain is irreducible and aperiodic; it is, however, not necessary to assume that S_0 started from the invariant distribution.

10.3 Statistical Modeling Based on Finite Markov Mixture Distributions

Researchers have found Markov mixture models increasingly useful in applied time series analysis.

10.3.1 The Basic Markov Switching Model

Assume that a time series $\{y_1, \ldots, y_T\}$ is observed as a single realization of a stochastic process $\{Y_1, \ldots, Y_T\}$. In the basic Markov switching model the time series $\{y_1, \ldots, y_T\}$ is assumed to be a realization of a stochastic process Y_t generated by a finite Markov mixture from a specific distribution family:

$$Y_t | S_t \sim \mathcal{T}(\boldsymbol{\theta}_{S_t}),$$

where S_t is an unobservable (hidden) K state ergodic Markov chain, and Y_t fulfills assumption **Y4**.

The basic Markov switching model found widespread applications in many practical areas including bioinformatics, biology, economics, finance, hydrology, marketing, medicine, and speech recognition. Various terminology became usual to denote models based on hidden Markov chains. The term Markov mixture models is preferred by biologists (Albert, 1991). Markov mixture models are usually called hidden Markov models in engineering applications (Zucchini and Guttorp, 1991; Thyer and Kuczera, 2000) and in speech recognition

(Levison et al., 1983; Rabiner, 1989). The terms Markov switching models or regime-switching models are preferred by economists who used Markov switching models to analyze stock market returns (Pagan and Schwert, 1990; Engel and Hamilton, 1990), interest rates (Ang and Bekaert, 2002) and asymmetries over the business cycle (Neftçi, 1984; Hamilton, 1989); see the monographs by Bhar and Hamori (2004), Krolzig (1997) and Kim and Nelson (1999) and Chapter 12 for further references and more details.

An interesting special case of the basic Markov switching model arises if $\{y_1, \ldots, y_T\}$ is a discrete-valued time series (MacDonald and Zucchini, 1997). Because one may choose Markov mixtures of any discrete distribution, it is possible to model many different types of discrete valued time series data, for example, binary time series by

$$\Pr(Y_t = 1|S_t) = \pi_{S_t}, \tag{10.27}$$

time series of bounded counts by a Markov mixture of binomial distributions,

$$Y_t|S_t \sim \text{BiNom}\left(n_t, \pi_{S_t}\right), \tag{10.28}$$

or time series of unbounded counts by a Markov mixture of Poisson distributions,

$$Y_t|S_t \sim \mathcal{P}\left(\mu_{S_t}\right); \tag{10.29}$$

see also Section 11.7. An important feature of applying Markov mixture models to discrete-valued time series is the ease with which autocorrelation is introduced, and the properties of the marginal distribution are easily analyzed.

Similarly, the basic Markov switching model could be applied to deal with autoregression in positive-valued time series (Lawrance and Lewis, 1985) simply by choosing the observation density $p(y_t|\boldsymbol{\theta})$ from any density on $\Re^+$, such as the exponential, the Gamma, or the Weibull distribution.

The basic Markov switching model has been generalized in several ways as outlined in the following subsections as well as in Chapter 12.

10.3.2 The Markov Switching Regression Model

An early attempt at introducing Markov switching models into econometrics in order to deal with time series data that depends on exogenous variables is the switching regression model of Goldfeld and Quandt (1973), which extends the switching regression model (Quandt, 1972) described earlier in Section 8.2. Whereas Quandt (1972) assumes that S_t is an i.i.d. random sequence, Goldfeld and Quandt (1973) allow explicitly for dependence between the states by modeling S_t as a two-state hidden Markov chain.

The general Markov switching regression model reads,

$$Y_t = \mathbf{x}_t\boldsymbol{\beta}_{S_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_t}\right), \tag{10.30}$$

where S_t is a hidden Markov chain and $\mathbf{x}_t$ is a row vector of explanatory variables including the constant (Lindgren, 1978; Cosslett and Lee, 1985). For discrete-valued explanatory variables, the Markov switching regression model will suffer from the same identifiability problems as the standard finite mixture of regression models studied in Subsection 8.2.2, a fact that has remained unnoted in the literature.

10.3.3 Nonergodic Markov Chains

In certain applications it makes sense to consider Markov switching models driven by a nonergodic Markov chain. An important example is a model driven by a Markov chain with transition matrix $\boldsymbol{\xi}$ defined in (10.6) which captures a single structural break or change-point. Assume that the Markov chain starts in $S_0 = 1$. The Markov chain will stay in state 1 for h periods; that is, $S_1 = \cdots = S_h = 1$ with probability ξ_{11}^h. Once state 2 is reached for the first time, the process remains there. An important aspect of this model is that the time of change-point occurrence is random.

A multiple change-point model with K change-points may be modeled through a Markov switching model with the following transition matrix (Chib, 1998),

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & 1-\xi_{11} & 0 & \cdots & 0 \\ 0 & \xi_{22} & 1-\xi_{22} & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ & & 0 & \xi_{K-1,K-1} & 1-\xi_{K-1,K-1} \\ & & & 0 & 1 \end{pmatrix}. \tag{10.31}$$

A more general Bayesian time series model of multiple structural changes in level, trend, and variance is studied in Wang and Zivot (2000). For a review of other methods of testing for the presence of unknown breakpoints in normal linear regression see Ploberger et al. (1989) and Andrews et al. (1996).

10.3.4 Relaxing the Assumptions of the Basic Markov Switching Model

The basic Markov switching model has been extended by many authors with the aim of formulating even more flexible models for a wide range of time series data.

Let $\{S_t\}_{t=0}^T$ be a finite-state Markov process with state space $\{1, \ldots, K\}$, and let $\{Y_t\}_{t=1}^T$ be a sequence of random variables with sampling space $\mathcal{Y}$. A general Markov switching model is obtained by specifying the density $p(\mathbf{S}, \mathbf{y}|\boldsymbol{\vartheta})$ of the joint distribution of $\mathbf{S} = \{S_t\}_{t=0}^T$ and $\mathbf{Y} = \{Y_t\}_{t=1}^T$, which is equal to:

$$p(\mathbf{S}, \mathbf{y}|\boldsymbol{\vartheta}) = p(S_0|\boldsymbol{\vartheta}) \prod_{t=1}^{T} p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta}) p(S_t|\mathbf{S}^{t-1}, \mathbf{y}^{t-1}, \boldsymbol{\vartheta}). \tag{10.32}$$

$p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ is the one-step ahead predictive density of the conditional distribution of Y_t, knowing the past realizations $\mathbf{y}^{t-1} = (y_1, \ldots, y_{t-1})$ of $\mathbf{Y}^{t-1}$ and knowing the states $\mathbf{S}^t = (S_0, \ldots, S_t)$. $p(S_t|\mathbf{y}^{t-1}, \mathbf{S}^{t-1}, \boldsymbol{\vartheta})$ is the density of the conditional distribution of S_t, knowing all past states $\mathbf{S}^{t-1} = (S_0, \ldots, S_{t-1})$ and the past realizations $\mathbf{y}^{t-1}$. The parameter $\boldsymbol{\vartheta}$ contains unknown model parameters such as the transition matrix $\boldsymbol{\xi}$, and other parameters indexing the densities $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ and $p(S_t|\mathbf{S}^{t-1}, \boldsymbol{\vartheta})$.

The basic Markov switching model, formulated in Subsection 10.2.1, results under rather strong assumptions concerning the densities $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ and $p(S_t|\mathbf{S}^{t-1}, \mathbf{y}^{t-1}, \boldsymbol{\vartheta})$. Under assumption **Y4**, the density $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ is not allowed to depend on past realizations $\mathbf{y}^{t-1}$ nor on the previous states of $\mathbf{S}^{t-1}$: $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta}) = p(y_t|\boldsymbol{\theta}_{S_t})$. Assumption **S4** implies that the conditional distribution $p(S_t|\mathbf{S}^{t-1}, \mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ is influenced by the state of S_{t-1}, only, and is independent of t. More general Markov switching models result by considering more general observation densities $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ or more general probability models of the hidden Markov chain.

More General Observation Densities

First of all, the conditional distribution of Y_t given S_t may be allowed to depend on past realizations $\mathbf{y}^{t-1} = (y_1, \ldots, y_{t-1})$ of $Y_1, \ldots, Y_{t-1}$, leading to assumption

Y3 Only the present value of S_t influences the density $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ and dependence on past values of S_t is not allowed:

$$p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta}) = p(y_t|\mathbf{y}^{t-1}, S_t, \boldsymbol{\vartheta}), \tag{10.33}$$

for $t = 1, \ldots, T$. Furthermore, $p(y_t|\mathbf{y}^{t-1}, S_t, \boldsymbol{\vartheta})$ is allowed to depend on exogenous variables $\boldsymbol{z}_t$.

The Markov switching regression model discussed in Subsection 10.3.2 results as that special case where $p(y_t|S_t, \boldsymbol{\vartheta})$ is independent of $\mathbf{y}^{t-1}$ while depending on exogenous variables $\boldsymbol{z}_t$. Further examples are the Markov switching autoregressive model suggested by McCulloch and Tsay (1994b), which is discussed in Section 12.2, and the Markov switching dynamic regression model, discussed in Section 12.3.

Assumption **Y3** is not fulfilled by the original Markov switching autoregressive model suggested by Hamilton (1989), which fulfills the more general condition

Y2 The present value of S_t, as well as a limited number of past values $S_{t-1}, \ldots, S_{t-p}$ influences the observation density $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$:

$$p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta}) = p(y_t|\mathbf{y}^{t-1}, S_t, \ldots, S_{t-p}, \boldsymbol{\vartheta}). \tag{10.34}$$

Assumption **Y2** is still too restrictive for switching ARMA models (Billio and Monfort, 1998) and switching GARCH models (Francq et al., 2001); see also Subsection 12.5.5. These models fulfill only the most general assumption

Y1 The observation density $p(y_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ depends on $\mathbf{y}^{t-1}$ and all past states of $\mathbf{S}^t$.

More General Models for the Hidden Markov Chain

The change-point model discussed in Subsection 10.3.3 shows that sensible Markov switching models result, when assumption **S4** is relaxed in the following way.

S3 S_t is a first-order homogeneous Markov chain with arbitrary transition matrix $\boldsymbol{\xi}$, which need not be irreducible or aperiodic, and starts from an arbitrary distribution $\mathbf{p}_0 = (p_{0,1}, \ldots, p_{0,K})$, where

$$p_{0,k} = \Pr(S_0 = k). \tag{10.35}$$

Furthermore it is possible to relax the assumption of homogeneity of the hidden Markov chain S_t as done in Subsection 12.6.1 for models with time-varying transition probabilities:

S2 S_t is a first-order inhomogeneous Markov chain, with the conditional distribution of S_t being independent of $\mathbf{y}^{t-1}$ and depending on the most recent value S_{t-1} and on some exogenous variables $\boldsymbol{z}_t$:

$$\Pr(S_t = k|\mathbf{S}^{t-1}, \mathbf{y}^{t-1}) = \Pr(S_t = k|S_{t-1}, \boldsymbol{z}_t), \qquad \forall k \in \{1, \ldots, K\}.$$

Some Markov switching models with time-varying transition matrices also allow for dependence of the transition matrix on previous realizations $\mathbf{y}^{t-1}$.

S1 S_t is a first-order Markov chain, and the conditional distribution of S_t depends on the history $\mathbf{y}^{t-1}$ of Y_t:

$$\Pr(S_t = k|\mathbf{S}^{t-1}, \mathbf{y}^{t-1}) = \Pr(S_t = k|S_{t-1}, \mathbf{y}^{t-1}), \qquad \forall k \in \{1, \ldots, K\},$$

for $t = 1, \ldots, T$.

The Initial Distribution of S_0

To complete the model specification for the process S_t, the distribution $\mathbf{p}_0$ needs to be specified. Under assumption **S4**, S_t starts from the ergodic probability distribution, hence $\mathbf{p}_0 = \boldsymbol{\eta}$. This assumption could be relaxed by assuming that S_t starts from an arbitrary discrete probability distribution $\mathbf{p}_0$, independent of $\boldsymbol{\xi}$. Note that the resulting Markov chain is no longer stationary.

The initial distribution $\mathbf{p}_0$ could either be a uniform distribution over $\{1, \ldots, K\}$ (Frühwirth-Schnatter, 2001b), or could be treated as an unknown parameter to be estimated from the data (Goldfeld and Quandt, 1973; Leroux and Puterman, 1992).

For certain reducible Markov chains it is sensible to assume that the starting value S_0 is a known value. Consider, for instance, the transition matrix (10.31), which captures structural breaks at unknown time points when starting with $S_0 = 1$.