# 13

# Switching State Space Models

## 13.1 State Space Modeling

As an introduction into the vast area of state space modeling, we start in Subsection 13.1.1 with the local level model which is a simple but characteristic example of the linear Gaussian state space form that is discussed in full generality in Section 13.1.2.

### 13.1.1 The Local Level Model with and Without Switching

In a local level model, a random process $\{Y_1, \dots, Y_T\}$ is generated by the following stochastic difference equation,

$$\mu_t = \mu_{t-1} + w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \tag{13.1}$$

$$Y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{13.2}$$

where all error terms $w_t$ and $\varepsilon_t$ are mutually independent and independent of $\mu_0$; see, for instance, Durbin and Koopman (2001) for an excellent introduction. The random process $\{Y_1, \dots, Y_T\}$ is assumed to be observable and the realizations are denoted by $\{y_1, \dots, y_T\}$. The distribution of $Y_t$ is allowed to depend on an unobservable latent variable, in this case the level $\mu_t$, which follows a random walk. Because $\mu_t$ is a hidden Markov process, this model is related to the hidden Markov chain model considered in Chapter 10; the latent process, however, does not live on a discrete state space, but on a continuous one.

The process $Y_t$ is nonstationary as long as the variance $\sigma_\mu^2$ is positive. There is a close relationship between the local level model and more classical time series models; see, for instance, Abraham and Ledolter (1986). By taking first differences, we obtain:

$$\Delta Y_t = w_t + \varepsilon_t - \varepsilon_{t-1}.$$

The lag 1 autocorrelation is given by

$$\rho_{\Delta Y_t}(1) = -\sigma_\varepsilon^2/(2\sigma_\varepsilon^2 + \sigma_\mu^2),$$

whereas higher autocorrelations are zero. Because this autocorrelation function is the same as that of an MA(1) process, the local level model has an ARIMA$(0,1,1)$ representation, where the MA(1) coefficient $\theta_1$ is constrained to the interval $[0,1]$ and results from equating the lag 1 autocorrelation in both models:

$$\rho_{\Delta Y_t}(1) = \frac{\theta_1}{1+\theta_1^2} \Rightarrow \theta_1 = \frac{1-\sqrt{1-4\rho_{\Delta Y_t}(1)^2}}{2\rho_{\Delta Y_t}(1)}.$$

The advantage of the state space form (13.1) and (13.2) as compared to the ARIMA$(0,1,1)$ representation is manifold as discussed extensively in the monograph of West and Harrison (1997). First, one may extract much more information from the observed time series as it is possible to estimate the level $\mu_t$ for each $t$ using the Kalman filter; see Section 13.3. Second, further components capturing seasonal patterns in the time series or trend behavior are easily added, as discussed in Subsection 13.2.1. Third, it is much easier to deal with time series irregularities such as outliers or structural breaks in the state space form.

Consider, for instance, a process generated by a local level model that is disrupted by occasional observation outliers. Based on the finite mixture approach to outlier modeling discussed in Section 7.2, the local level model may be modified in the following way,

$$\begin{aligned} \mu_t &= \mu_{t-1} + w_t, \qquad & w_t &\sim \mathcal{N}\left(0, \sigma_\mu^2\right), \\ Y_t &= \mu_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \eta_1 \mathcal{N}\left(0, \sigma_{\varepsilon,1}^2\right) + \eta_2 \mathcal{N}\left(0, \sigma_{\varepsilon,2}^2\right). \end{aligned}$$

After introducing an i.i.d. binary indicator $S_t$ with $\Pr(S_t = 1) = \eta_1$ as in earlier chapters, conditional on knowing $S_t$, this model is a local linear model as defined in (13.1) and (13.2), however, the observation variance $\sigma_\varepsilon^2$ is switching between two values:

$$\begin{aligned} \mu_t &= \mu_{t-1} + w_t, \qquad & w_t &\sim \mathcal{N}\left(0, \sigma_\mu^2\right), \\ Y_t &= \mu_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathcal{N}\left(0, \sigma_{\varepsilon,S_t}^2\right). \end{aligned}$$

This is a first example of a switching state space model, which is commonly applied to deal with outliers in time series; see Subsection 13.2.3 for a more detailed discussion. Another useful switching state space model results if $S_t$ follows a hidden Markov chain as in Chapter 10, rather than an i.i.d. process, because this introduces conditional heteroscedasticity in the error term $\varepsilon_t$; see also Subsection 13.2.2.

Apart from the observation variance, other model parameters may be switching in a state space model. Consider, for instance, the variance $\sigma_\mu^2$ in the local level model, which determines how much $\mu_t$ changes over time. The smaller $\sigma^2_\mu$, the less flexibility of $\mu_t$ is allowed a priori. To distinguish between periods of smaller and greater variability of $\mu_t$ it may be assumed that the variance $\sigma^2_\mu$ switches between two values:

$$\mu_t = \mu_{t-1} + w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma^2_{\mu,S_t}\right).$$

A special case of this model is one where $\sigma^2_{\mu,1}$ is 0, whereas $\sigma^2_{\mu,2} > 0$. Such a model allows for an occasional level shift:

$$\mu_t = \begin{cases} \mu_{t-1}, & S_t = 1 \\ \mu_{t-1} + w_t, \quad w_t \sim \mathcal{N}\left(0, \sigma^2_\mu\right), & S_t = 2. \end{cases} \tag{13.3}$$

Finally, two independent indicators $S^1_t$ and $S^2_t$ may be introduced to combine heteroscedasticity in $w_t$ with observation outliers:

$$\begin{aligned} \mu_t &= \mu_{t-1} + w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma^2_{\mu,S^1_t}\right), \\ Y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S^2_t}\right). \end{aligned}$$

### 13.1.2 The Linear Gaussian State Space Form

The local level model introduced in the previous subsection is a special case of a linear Gaussian state space model, which is a dynamic stochastic system defined in the following way,

$$\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}_t\right), \tag{13.4}$$

$$\mathbf{Y}_t = \mathbf{H}_t\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{R}_t\right), \tag{13.5}$$

where $t = 1, \ldots, T$. The key variables in these formulations are the state variable $\mathbf{x}_t$ and the observation variable $\mathbf{Y}_t$.

The state variable $\mathbf{x}_t$ is a latent $d$-dimensional random vector, which is observed only indirectly through the effect it has on the distribution of $\mathbf{Y}_t$. The transition equation (13.4), also called the state equation, specifies for each $t \geq 1$ how $\mathbf{x}_t$ is generated from the previous state variable $\mathbf{x}_{t-1}$. The linear relationship between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ which depends on the $(d \times d)$ matrix $\mathbf{F}_t$ is disturbed by a zero-mean error $\mathbf{w}_t$ following a normal distribution with variance–covariance matrix $\mathbf{Q}_t$. To complete the model formulation, the distribution of $\mathbf{x}_0$ is specified as $\mathbf{x}_0 \sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{0|0}, \mathbf{P}_{0|0}\right)$.

The observation variable $\mathbf{Y}_t$ is a random vector of dimension $r$, which is assumed to be observable for all time points $t = 1, \ldots, T$. A single realization of this process is denoted by $\{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$. The dimension of $\mathbf{Y}_t$ may be smaller, larger or equal to the dimension of $\mathbf{x}_t$. For a scalar observation variable with $r = 1$ we write $Y_t$ and denote the observed time series by $\{y_1, \ldots, y_T\}$. The observation equation (13.5), also called the measurement equation, specifies how the distribution of $\mathbf{Y}_t$ is influenced by the state variable $\mathbf{x}_t$. The linear relationship between $\mathbf{Y}_t$ and $\mathbf{x}_t$ which depends on the $(r \times d)$ matrix $\mathbf{H}_t$ is disturbed by a zero-mean random observation error $\boldsymbol{\varepsilon}_t$ following a normal distribution with variance–covariance matrix $\mathbf{R}_t$, which reduces to a scalar variance for $r = 1$.

Often the matrices $\mathbf{F}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$, and $\mathbf{R}_t$ emerge from putting a specific time series model into a state space form; see Subsection 13.2.1. The local level model introduced in Subsection 13.1.1, for instance, results with $\mathbf{x}_t = \mu_t$, $\mathbf{F}_t = \mathbf{H}_t = 1$, $\mathbf{Q}_t = \sigma^2_\mu$, and $\mathbf{R}_t = \sigma^2_\varepsilon$. The matrices $\mathbf{F}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$, and $\mathbf{R}_t$ need not depend on time, in which case the notation $\mathbf{F}$, $\mathbf{H}$, $\mathbf{Q}$, and $\mathbf{R}$ will be used.

Some elements of the matrices $\mathbf{F}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$, and $\mathbf{R}_t$ may depend on unknown model parameters $\boldsymbol{\vartheta}$, such as for the local level model, where $\boldsymbol{\vartheta} = (\sigma^2_\mu, \sigma^2_\varepsilon)$. The notations $\mathbf{F}_t(\boldsymbol{\vartheta})$, $\mathbf{H}_t(\boldsymbol{\vartheta})$, $\mathbf{Q}_t(\boldsymbol{\vartheta})$, and $\mathbf{R}_t(\boldsymbol{\vartheta})$ are used whenever it is necessary to make this dependence explicit. Identification becomes an important issue whenever part of the system matrices $\mathbf{F}_t(\boldsymbol{\vartheta})$ and $\mathbf{H}_t(\boldsymbol{\vartheta})$ are unknown; see Hannan and Deistler (1988) for an extensive treatment of this issue.

Further assumptions are necessary to complete the model definition. Most important, $\mathbf{w}_t$ is uncorrelated with $\mathbf{x}_{t-1}$ for all $t$:

$$\mathrm{E}\left(\mathbf{w}_t\mathbf{x}_{t-1}^{'}\right) = \mathbf{0}, \qquad t = 1, \ldots, T.$$

Second, the observation error $\boldsymbol{\varepsilon}_t$ as well as $\mathbf{w}_t$ is uncorrelated over time:

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s^{'}\right) = \mathbf{0}, \qquad \mathrm{E}\left(\mathbf{w}_t\mathbf{w}_s^{'}\right) = \mathbf{0}, \qquad \forall s, t \in \{1, \ldots, T\}, t \neq s.$$

Finally, the two error sequences $\boldsymbol{\varepsilon}_t$ and $\mathbf{w}_s$ are uncorrelated for all $t, s$:

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_t\mathbf{w}_s^{'}\right) = \mathbf{0}, \qquad \forall s, t \in \{1, \ldots, T\}.$$

On various occasions, it is useful to introduce additional terms that influence the mean of the transition as well as the observation equation:

$$\mathbf{x}_t = \mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{G}_t\mathbf{u}_t + \mathbf{w}_t, \tag{13.6}$$

$$\mathbf{Y}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{A}_t\mathbf{z}_t + \boldsymbol{\varepsilon}_t. \tag{13.7}$$

In (13.6), $\mathbf{u}_t$ is a vector of dimension $n$, which may be smaller, larger, or equal to the dimension of $\mathbf{x}_t$. In engineering applications $\mathbf{u}_t$ often is a controllable input vector (Anderson and Moore, 1979). In econometric problems $\mathbf{u}_t$ often is a vector of $n$ exogenous variables being observable at time $t$. The expected value of $\mathbf{x}_t$ given $\mathbf{u}_t$ and $\mathbf{x}_{t-1}$ is a linear function in $\mathbf{u}_t$, depending on the $(d \times n)$ matrix $\mathbf{G}_t$. In (13.7), $\mathbf{z}_t$ is a vector of $m$ variables being observable at time $t$, which could be exogenous variables or past values of $\mathbf{Y}_t$. The expected value of $\mathbf{Y}_t$ given $\mathbf{x}_t$ and $\mathbf{z}_t$ is a linear function in $\mathbf{z}_t$, depending on the $(r \times m)$ matrix $\mathbf{A}_t$. For a further review of various aspects of state space modeling we refer to the monographs of Aoki (1990), Harvey (1993), West and Harrison (1997), and Durbin and Koopman (2001).

Originally, the state space model was developed by Kalman (1960, 1961) in aerospace research for tracking some target such as an aircraft. In this application the transition equation is derived from physical laws describing the motion of the target, whereas the observation vector measures properties of this target that are observable through some device such as a radar, subject to some measurement error; see also related tracking problems in high-energy physics (Frühwirth, 1987). Due to their flexibility and generality state space models found applications in many research areas in engineering such as hydrology (Schnatter et al., 1987) and speech recognition (Juang and Rabiner, 1985; Rabiner and Juang, 1986), just to mention two; see Anderson and Moore (1979) for further references.

The application of state space models in the econometric literature started in the 1970s with the time-varying coefficient model; see the review of Nicholls and Pagan (1985). In the 1980s, it was recognized that econometric models that rely on unobservable quantities could be cast into a state space form, and state space models found wide applications in economics and finance, for instance, to estimate the ex ante real interest rate (Fama and Gibbons, 1982), unobserved expected inflation (Burmeister et al., 1986), or the potential real GDP (Kuttner, 1994); see Granger and Teräsvirta (1993) and Kim and Nelson (1999) for further applications of state space models in econometrics.

### 13.1.3 Multiprocess Models

The simplest way of introducing a latent discrete indicator into the linear Gaussian state space form is multiprocess models. A multiprocess model is a collection of $K$ state space models, indexed by a hidden random indicator $S$ taking values in a discrete space $\{1, \dots, K\}$. Conditional on knowing the state of $S$, the model for $\mathbf{Y}_t$ is a linear Gaussian state space form:

$$\mathbf{x}_t = \mathbf{F}_t^{[S]}\mathbf{x}_{t-1} + \mathbf{G}_t^{[S]}\mathbf{u}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}_t^{[S]}\right), \tag{13.8}$$

$$\mathbf{Y}_t = \mathbf{H}_t^{[S]}\mathbf{x}_t + \mathbf{A}_t^{[S]}\mathbf{z}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{R}_t^{[S]}\right). \tag{13.9}$$

Multiprocess models were well known in the control engineering literature for many years (see, for instance, Magill, 1965) before they were introduced into the statistics literature by Harrison and Stevens (1976). Multiprocess models were applied to forecasting multiple time series (Schnatter et al., 1987), to deal with unobserved heterogeneity in longitudinal studies (Gamerman and Smith, 1996), or to cluster time series in panel data (Frühwirth-Schnatter and Kaufmann, 2006b).

### 13.1.4 Switching Linear Gaussian State Space Models

The basic idea of a switching state space model is that a priori no single model is expected to hold for all time points $t$, rather the possibility that different models hold at different times points is explicitly recognized by modeling the hidden model indicator $S_t$ as being dynamic over time.

A switching linear Gaussian state space model is based on the state space form introduced in Subsection 13.1.2, however, some (or all) system matrices are driven by a hidden model indicator $S_t$:

$$\mathbf{x}_t = \mathbf{F}_t^{[S_t]}\mathbf{x}_{t-1} + \mathbf{G}_t^{[S_t]}\mathbf{u}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}_t^{[S_t]}\right), \tag{13.10}$$

$$\mathbf{Y}_t = \mathbf{H}_t^{[S_t]}\mathbf{x}_t + \mathbf{A}_t^{[S_t]}\mathbf{z}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{R}_t^{[S_t]}\right). \tag{13.11}$$

$\{S_t, t = 1, \dots, T\}$ is a sequence of random variables, allowed to take values in the discrete space $\{1, \dots, K\}$. The degenerate case $S_t \equiv S_{t-1} \equiv S$ reduces to the multiprocess model introduced in Subsection 13.1.3.

To complete the model specification, some probabilistic structure has to be imposed on $S_t$. We distinguish two cases of switching state space models, namely finite mixtures of state space models, if $S_t$ is an i.i.d. sequence with probability distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$, and Markov switching state space models, if $S_t$ is a hidden Markov chain with transition matrix $\boldsymbol{\xi}$ as introduced in Chapter 10. The first structure may be regarded as a special case of the second structure with restricted transition matrix; see Subsection 10.2.6. Finite mixtures of state space models are sometimes called "multi process models" (Harrison and Stevens, 1976; Smith and West, 1983), whereas Markov switching state space models are sometimes called "state space models with regime switching" (Kim and Nelson, 1999).

The engineering literature has seen several pioneering works on switching state space models since the 1960s. For target tracking problems, for instance, Nahi (1969) assumes a nonzero probability that any observation consists of noise only, leading to a state space model where $\mathbf{H}_t$ is switching between a zero and a nonzero value; see Bar-Shalom and Tse (1975) and Shumway and Stoffer (1991) for a related application. In control engineering research Ackerson and Fu (1970) consider a linear Gaussian state space model, where the covariance matrices in the transition and in the observation equation are allowed to depend on a hidden Markov chain.

Harrison and Stevens (1976) introduced finite mixtures of linear Gaussian state space models into the statistics literature; further applications are found in medicine (Smith and West, 1983; Gordon and Smith, 1990), speech recognition (Juang and Rabiner, 1985; Rabiner and Juang, 1986), and hydrology (Schnatter, 1988b). The Markov switching linear Gaussian state space model became popular in econometrics through the work of Kim (1993a, 1993b, 1994) and Shephard (1994); see also the monograph of Kim and Nelson (1999) for further applications and references.

### 13.1.5 The General State Space Form

A useful way of thinking of a state space model is in terms of a hierarchical model, where on a first level the model specifies the conditional distribution

$p(\mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of the process $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ given the whole state process $\mathbf{x}_1, \ldots, \mathbf{x}_T$. On a second level the model characterizes the distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ of the state process. The following structure is characteristic of a state space model. The random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_T$ are independent of each other given the state process $\mathbf{x}_1, \ldots, \mathbf{x}_T$ and $\mathbf{Y}_t$ is independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$ given $\mathbf{x}_t$:

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{x}_1, \ldots, \mathbf{x}_T) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t). \tag{13.12}$$

The state variable $\mathbf{x}_t$ is a first-order hidden Markov process, hence independent of $\mathbf{x}_1, \ldots, \mathbf{x}_{t-2}$ given $\mathbf{x}_{t-1}$:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}). \tag{13.13}$$

Thus to define a state space model, one could directly specify for each $t = 1, \ldots, T$ the observation density $p(\mathbf{y}_t|\mathbf{x}_t)$ and the transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Because these densities are in principle arbitrary, the hierarchical formulation is very useful, as it allows us to introduce nonlinearities in the relationship between $\mathbf{y}_t$ and $\mathbf{x}_t$ and $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ as well as nonnormality by densities that are intrinsically nonnormal.

For a linear Gaussian state space model the observation and the transition density evidently are given by

$$\begin{aligned} \mathbf{Y}_t|\mathbf{x}_t &\sim \mathcal{N}_r\left(\mathbf{H}_t\mathbf{x}_t + \mathbf{A}_t\mathbf{z}_t, \mathbf{R}_t\right), \\ \mathbf{x}_t|\mathbf{x}_{t-1} &\sim \mathcal{N}_d\left(\mathbf{F}_t\mathbf{x}_{t-1} + \mathbf{G}_t\mathbf{u}_t, \mathbf{Q}_t\right). \end{aligned}$$

For a switching state space model the distributions (13.12) and (13.13) are formulated conditional on knowing the hidden indicators $\mathbf{S} = (S_0, S_1, \ldots, S_T)$, whereas a third level is added by describing the probability law for $\mathbf{S}$:

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_T|\mathbf{x}_1, \ldots, \mathbf{x}_T, \mathbf{S}) &= \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t, S_t), \\ p(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{S}) &= \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, S_t), \\ p(\mathbf{S}) &= \prod_{t=1}^{T} p(S_t|S_{t-1}). \end{aligned}$$

As in Chapter 10, the transition density as well as the observation density is assumed to depend on $S_t$ only.

## 13.2 Nonlinear Time Series Analysis Based on Switching State Space Models

### 13.2.1 ARMA Models with and Without Switching

The linear Gaussian state space form introduced in Subsection 13.1.2 subsumes many models that are popular in time series analysis, including regression models and ARMA models (Harvey, 1989; Shumway and Stoffer, 2000; Durbin and Koopman, 2001). Consider the ARMA$(p,q)$ process,

$$\boldsymbol{\delta}(L)(Y_t-\mu)=\boldsymbol{\theta}(L)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0,\sigma_\varepsilon^2\right),$$

where $L$ is the lag operator, $\boldsymbol{\delta}(L)=1-\delta_1 L-\cdots-\delta_p L^p$, and $\boldsymbol{\theta}(L)=1-\theta_1 L-\cdots-\theta_q L^q$, with $\delta_1,\ldots,\delta_p$ being the AR coefficients, and $\theta_1,\ldots,\theta_q$ being the MA coefficients. This model possesses for $q=p-1$ the following state space representation with $\mathbf{x}_t\in\Re^p$,

$$Y_t=\mu+\mathbf{H}(\boldsymbol{\theta})\mathbf{x}_t, \tag{13.14}$$

$$\mathbf{x}_t=\mathbf{F}(\boldsymbol{\delta})\mathbf{x}_{t-1}+\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}\varepsilon_t, \qquad \varepsilon_t\sim\mathcal{N}\left(0,\sigma_\varepsilon^2\right), \tag{13.15}$$

with

$$\mathbf{F}(\boldsymbol{\delta})=\begin{pmatrix}\delta_1\ldots\delta_{p-1} & \delta_p\\ \mathbf{I}_{p-1} & \mathbf{0}_{(p-1)\times 1}\end{pmatrix}, \qquad \mathbf{H}(\boldsymbol{\theta})=\begin{pmatrix}1 & -\theta_1 & \cdots & -\theta_q\end{pmatrix}. \tag{13.16}$$

The same state space representation could be used if $q<p-1$ after adding $(p-1-q)$ MA coefficients equal to 0: $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_q,0,\ldots,0)$. If $q\geq p$, a similar state space representation with $\mathbf{x}_t\in\Re^{q+1}$ could be used, with the matrix $\mathbf{F}(\boldsymbol{\delta})$ in (13.15) being defined after adding $(q-p+1)$ AR coefficients equal to 0: $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_p,0,\ldots,0)$.

In complex modeling situations it is often easier to work with the state space representation, in particular when dealing with outliers, missing data, interventions, mixed-effects, and structural changes (Kohn and Ansley, 1986; Harvey et al., 1998).

This is, for instance, the case if a hidden Markov chain $S_t$ is introduced into an ARMA model, one example being the switching ARMA model (Billio and Monfort, 1998; Billio et al., 1999), for which the one-step ahead predictive density depends on the whole history of $S_t$. This long-range dependence disappears conditional on the latent variables $\mathbf{x}_t$ and $S_t$, if an ARMA$(p,q)$ process with switching mean is represented by the following switching state space model,

$$Y_t=\mu_{S_t}+\mathbf{H}(\boldsymbol{\theta})\mathbf{x}_t,$$

with the state equation being the same as in (13.15). This facilitates statistical inference in Section 13.3 and 13.4.

A similar result holds for the Markov switching autoregressive model of Lam (1990), defined earlier in (12.11), which has the following state space form,

$$Y_t = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ \mathbf{x}_t \end{pmatrix},$$

$$\begin{pmatrix} \mu_t \\ \mathbf{x}_t \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{F}(\boldsymbol{\delta}) \end{pmatrix} \begin{pmatrix} \mu_{t-1} \\ \mathbf{x}_{t-1} \end{pmatrix} + \begin{pmatrix} \beta_{S_t} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \varepsilon_t,$$

with $\mathbf{x}_t$ and $\mathbf{F}(\boldsymbol{\delta})$ being the same as in (13.15).

### 13.2.2 Unobserved Component Time Series Models

The state space approach is also useful for decomposing a time series into unobserved components such as trend, cycles, seasonal, and irregular components (Harvey, 1989). A simple example of such a model is the local level model discussed in Subsection 13.1.1; a more flexible one is the basic structural model (Harvey and Todd, 1983):

$$\begin{aligned}
\mu_t &= \mu_{t-1} + \beta_{t-1} + w_{t,1}, \qquad w_{t,1} \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \\
\beta_t &= \beta_{t-1} + w_{t,2}, \qquad w_{t,2} \sim \mathcal{N}\left(0, \sigma_\beta^2\right), \\
\gamma_t &= -\sum_{j=1}^{s-1} \gamma_{t-j} + w_{t,3}, \qquad w_{t,3} \sim \mathcal{N}\left(0, \sigma_\gamma^2\right), \\
Y_t &= \mu_t + \gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \qquad (13.17)
\end{aligned}$$

$\mu_t$ is the slowly varying trend of the time series, $\gamma_t$ is a periodic seasonal component, and $\varepsilon_t$ is a random disturbance term. If no seasonal component is present in (13.17) then the resulting model is called the local linear trend model.

Decomposing a time series into a stochastic and stationary component may lead to identification problems (Nelson, 1988). The local level model discussed in Subsection 13.1.1, for instance, is not identified if the two noise terms $\varepsilon_t$ and $w_t$ are allowed to be correlated. For this reason, it is assumed in the basic structural model that all error terms are uncorrelated.

Unobserved component models found numerous applications in economics and have been extended in several ways by including a hidden indicator. Many applications of this model typically are based on the assumption that the error terms in the state and in the observation equation are homoscedastic. Heteroscedasticity may be caused by outliers as discussed in Subsection 13.2.3.

In addition, it is reasonable to assume there exists some kind of conditional heteroscedasticity in that errors with large variances tend to be followed by errors with large variances and similarly errors with small variances tend to be followed by errors with small variances.

To capture heteroscedasticity, Harvey et al. (1992) consider unobserved component models with ARCH disturbances both in the transition as well as in the observation equation. As an alternative, Kim (1993b) introduced unobserved component time series models with Markov switching heteroscedasticity, by assuming that the variances depend on a hidden Markov chain:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}_t^{[S_t]}\right), \\ Y_t &= \mathbf{H}\mathbf{x}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_t}\right). \end{aligned}$$

Applications of this model include modeling the link between inflation rates and inflation uncertainty (Kim, 1993b) and analyzing the U.S. stock market with focus on the 1987 crash (Kim and Kim, 1996).

Alternatively, hidden indicators have been introduced into the structural part of unobserved component models. Whittaker and Frühwirth-Schnatter (1994) define a dynamic change-point model, where in a local level model, a random walk drift is added after a structural break:

$$\begin{aligned} \mu_t &= \mu_{t-1} + (S_t - 1)\beta_{t-1} + w_{t,1}, \qquad w_{t,1} \sim \mathcal{N}\left(0, \sigma^2_\mu\right), \\ \beta_t &= \beta_{t-1} + w_{t,2}, \qquad w_{t,2} \sim \mathcal{N}\left(0, \sigma^2_\beta\right), \end{aligned}$$

where $S_t$ is allowed a one-time change between state 1 and 2 at an unknown change-point $\tau$.

To capture different growth behavior in boom and recession, Luginbuhl and de Vos (1999) model the log gross domestic product by a switching local linear trend model. Two different drift components $\alpha_t$ and $\beta_t$ are assumed to be present, each of which follows a random walk, but only one of them contributes to the trend:

$$\begin{aligned} \mu_t &= \mu_{t-1} + (1 - S_t)\alpha_{t-1} + S_t\beta_{t-1} + w_{t,1}, \\ \alpha_t &= \alpha_{t-1} + w_{t,2}, \\ \beta_t &= \beta_{t-1} + w_{t,3}, \\ Y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right). \end{aligned}$$

The indicator $S_t$, selecting one of the two trend components, is assumed to follow a hidden Markov chain with state space $\{0, 1\}$.

### 13.2.3 Capturing Sudden Changes in Time Series

Detecting sudden changes, outliers, and level shifts is an important aspect of practical time series analysis, often called intervention analysis (Tsay, 1988).

Many authors generalized the linear Gaussian state space model with the aim of establishing recursive estimation procedures that are robust to outliers (Masreliez, 1975; Masreliez and Martin, 1975; West, 1981, 1984; Tsai and Kurz, 1983; Peña and Guttman, 1988; Meinhold and Singpurwalla, 1989). Peña and Guttman (1988) generalized the scale-contaminated model (Tukey, 1960; Box and Tiao, 1968), already discussed in Subsection 7.2.1, to the framework of robust linear Gaussian state space models with univariate observation vector $Y_t$, by assuming that the noise $\varepsilon_t$ in the observation equation (13.5) follows a mixture of two normal distributions with mean zero, but different variances:

$$\varepsilon_t \sim (1-\eta_2)\mathcal{N}\left(0,\sigma_\varepsilon^2\right) + \eta_2\mathcal{N}\left(0,k\sigma_\varepsilon^2\right),$$

where typically $\eta_2$ is a small fraction of outliers, whereas $k >> 1$. For estimation, however, it is useful to view such a robust state space model as a switching Gaussian state space model, where the distribution of the observation noise is driven by a hidden i.i.d. sequence $S_t$:

$$\sigma^2_{\varepsilon,S_t} = \begin{cases} \sigma_\varepsilon^2, & S_t = 1, \\ k\sigma_\varepsilon^2, & S_t = 2, \end{cases}$$

with probability $\Pr(S_t = 2) = \eta_2$.

In Meinhold and Singpurwalla (1989) robustness is achieved by assuming that both $\mathbf{w}_t$ and $\varepsilon_t$ have a marginal $t$-distribution of differing degree of freedom $\nu_1$ and $\nu_2$, which may be written as

$$\begin{aligned} \mathbf{w}_t &\sim \mathcal{N}_d\left(\mathbf{0},\mathbf{Q}/\omega_t^1\right), \qquad & \omega_t^1 &\sim \mathcal{G}\left(\nu_1/2,\nu_1/2\right), \\ \varepsilon_t &\sim \mathcal{N}\left(0,\sigma_\varepsilon^2/\omega_t^2\right), \qquad & \omega_t^2 &\sim \mathcal{G}\left(\nu_2/2,\nu_2/2\right). \end{aligned}$$

A combination of these two robust state space models appears in Godsill and Rayner (1998) for the reconstruction of signals that are degraded by an impulsive noise:

$$Y_t = \mathbf{x}_t + (S_t - 1)\upsilon_t, \qquad \upsilon_t \sim \mathcal{N}\left(0,\sigma_\upsilon^2/\omega_t\right), \qquad \omega_t \sim \mathcal{G}\left(\nu/2,\nu/2\right),$$

where $S_t$ is a hidden Markov chain taking the value 1, if no noise is present, and 2 otherwise. $\mathbf{x}_t$ is an AR($p$) process modeled through a state space model as in (13.15).

A more general model, where outliers may be observational as well as innovational is considered in Godsill (1997) in the context of reconstructing acoustically recorded signals, such as speech and music. The statistical model is an ARMA($p,q$) process observed with noise, which possesses the following state space representation with observation equation,

$$Y_t = \mu + \mathbf{H}(\boldsymbol{\theta})\mathbf{x}_t + \upsilon_t,$$

with $\mathbf{x}_t$ and $\mathbf{H}(\boldsymbol{\theta})$ being the same as in (13.15) and (13.16). Both $\upsilon_t$ as well as the error term $\varepsilon_t$ appearing in (13.15) are assumed to follow a mixture of a normal and a $t$-distribution:

$$\begin{aligned}
&\upsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\upsilon,S_t^1}\right), \\
&\quad \sigma^2_{\upsilon,S_t^1} = (2 - S_t^1)\sigma^2_\upsilon + (S_t^1 - 1)\sigma^2_\upsilon/\omega_t^1, \qquad \omega_t^1 \sim \mathcal{G}\left(\nu_1/2, \nu_1/2\right), \\
&\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_t^2}\right), \\
&\quad \sigma^2_{\varepsilon,S_t^2} = (2 - S_t^2)\sigma^2_\varepsilon + (S_t^2 - 1)\sigma^2_\varepsilon/\omega_t^2, \qquad \omega_t^2 \sim \mathcal{G}\left(\nu_2/2, \nu_2/2\right).
\end{aligned}$$

$S_t^1$ and $S_t^2$ are two independent two-state hidden Markov chains with unknown transition matrices $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$.

Another useful model to deal with structural or innovation outliers is the random level shift time series model (Chen and Tiao, 1990; McCulloch and Tsay, 1993):

$$\begin{aligned}
&Y_t = \mu_t + Z_t, \\
&\mu_t = \mu_{t-1} + (S_t - 1)\beta_t, \qquad \beta_t \sim \mathcal{N}\left(0, k\sigma^2_\varepsilon\right), \\
&\boldsymbol{\delta}(L)Z_t = \boldsymbol{\theta}(L)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right),
\end{aligned}$$

where $S_t$ is a two-state hidden i.i.d. indicator with $S_t = 2$ corresponding to a shift that occurs a priori with probability $\Pr(S_t = 2) = \eta_2$. If $\eta_2 = 1$, then the level changes all the time and the model is related to the local trend model (13.17). Gerlach and Kohn (2000) show how intervention analysis may be treated through a switching state space model including both a hidden Markov indicator as well as a second i.i.d. indicator to deal with outliers.

For any of these models traditional likelihood estimation is rather involved. The Bayesian framework discussed in Section 13.4 offers the possibility of locating the position and the size of outlier and shifts simultaneously with parameter estimation.

### 13.2.4 Switching Dynamic Factor Models

Dynamic factor models, in which a large number of observed time series are assumed to be influenced by a common unobserved component, are a special case of a state space model which found various applications in economics, for instance, to estimate wage rates (Engle and Watson, 1981) and to analyze economic indicators that move together (Stock and Watson, 2002).

Diebold and Rudebusch (1996) combine the dynamic factor model with the Markov switching model, one example being the following model,

$$\begin{aligned}
&\triangle \mathbf{Y}_t = \boldsymbol{\beta} + \boldsymbol{\lambda} f_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \\
&\boldsymbol{\delta}(L)(f_t - \mu_{S_t}) = w_t, \qquad w_t \sim \mathcal{N}\left(0, 1\right),
\end{aligned}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix, and $w_t$ and $\boldsymbol{\varepsilon}_t$ are pairwise independent. $f_t$ is the latent dynamic factor, $\boldsymbol{\beta}$ and the factor loadings $\boldsymbol{\lambda}$ are unknown parameters. Diebold and Rudebusch (1996) extended this model by considering more general structures for the error process $\boldsymbol{\varepsilon}_t$ such as a VAR model. Kim and

Nelson (1998) generalize this model by introducing time-varying transition matrices.

Application appeared mainly in business cycle analysis (Kim and Nelson, 1998, 2001; Kaufmann, 2000).

### 13.2.5 Switching State Space Models as a Semi-Parametric Smoothing Device

State space models are a useful device for smoothing and interpolating time series (Wecker and Ansley, 1983; Kohn and Ansley, 1987) which are closely related to semiparametric optimal smoothing methods based on the roughness penalty approach.

Kitagawa (1981), for instance, considers the following smoothness prior approach for smoothing nonstationary time series,

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \\ \mu_t - 2\mu_{t-1} + \mu_{t-2} &= w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \end{aligned} \tag{13.18}$$

which is closely related to basic structural model (13.17) without a seasonal component $\gamma_t$ and has a very simple state space form. A model that is similar to (13.17) was introduced by Kitagawa and Gersch (1984) for smoothing time series with trends and seasonal components.

Posterior mode estimation for model (13.18) under diffuse priors on $\mu_{-1}$ and $\mu_0$ corresponds to minimizing the penalized least square criterion

$$\sum_{t=1}^{T}(y_t - \mu_t)^2 + \lambda \sum_{t=3}^{T}(\mu_t - 2\mu_{t-1} + \mu_{t-2})^2, \tag{13.19}$$

with respect to $\mu_1, \dots, \mu_T$, where the smoothness parameter $\lambda$ is related to the variances of the error terms through $\lambda = \sigma_\varepsilon^2/\sigma_\mu^2$ (Fahrmeir and Knorr-Held, 2000). If in (13.18), the fixed variance $\sigma_\mu^2$ is substituted by the switching variance $\sigma_{\mu,S_t}^2$, then the smoothness parameter itself depends on the hidden Markov chain $S_t$: $\lambda_t = \sigma_\varepsilon^2/\sigma_{\mu,S_t}^2$. In this respect, switching state space models with heteroscedastic variances $\sigma_{\mu,S_t}^2$ may be seen as a device for smoothing time series where the smoothness parameter changes over time.

## 13.3 Filtering for Switching Linear Gaussian State Space Models

Filtering aims at deriving the posterior density $p(\mathbf{x}_t|\mathbf{y}^t, \boldsymbol{\vartheta})$ of $\mathbf{x}_t$ given observations $\mathbf{y}^t = (\mathbf{y}_1, \dots, \mathbf{y}_t)$ up to $t$ in an efficient manner for a fixed model parameter $\boldsymbol{\vartheta}$. To keep notation simple, dependence on $\boldsymbol{\vartheta}$ is not made explicit.

### 13.3.1 The Filtering Problem

Regrettably, the posterior density $p(\mathbf{x}_t|\mathbf{y}^t)$ is of closed form only for very restricted state space models with the linear Gaussian state space model being the most prominent one. For this model class, the posterior density $p(\mathbf{x}_t|\mathbf{y}^t)$ is a normal distribution, where the first two moments are given by the Kalman filter (Kalman, 1960, 1961); see also *Algorithm 13.1* below.

Long before the Bayesian community became aware of the Kalman filter, the importance of the Bayesian approach for solving the filtering problem was realized in the engineering literature (Magill, 1965; Alspach and Sorenson, 1972). As pointed out by Alspach and Sorenson (1972, p.439) regarding $p(\mathbf{x}_t|\mathbf{y}^t)$,

> *If this posterior density function were known, an estimate of the state for any performance criterion could be determined.*

Also for a nonlinear non-Gaussian state space model the filter problem is solved by recursions similar in structure, but not in complexity, to the Kalman filter. Let $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ be the posterior density of the state $\mathbf{x}_{t-1}$ given information up to $t-1$. The first part of the filtering step is to propagate this information into the future, by deriving the density $p(\mathbf{x}_t|\mathbf{y}^{t-1})$ which may be obtained from integrating the density $p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ with respect to $\mathbf{x}_{t-1}$. By assumption (13.13), the propagation step reads:

$$p(\mathbf{x}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})d\mathbf{x}_{t-1}.$$

Once an observation $\mathbf{y}_t$ is available, Bayes' theorem plays a crucial role in finding a coherent way of combining information propagated from the past with the information contained in $\mathbf{y}_t$. The updated posterior density $p(\mathbf{x}_t|\mathbf{y}^t)$ is obtained from Bayes' theorem as

$$p(\mathbf{x}_t|\mathbf{y}^t) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1})}{p(\mathbf{y}_t|\mathbf{y}^{t-1})},$$

with the normalizing constant being identical to the one-step ahead predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1})$:

$$p(\mathbf{y}_t|\mathbf{y}^{t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1})d\mathbf{x}_t.$$

### 13.3.2 Bayesian Inference for a General Linear Regression Model

It is useful to discuss the filtering problem first for a multivariate regression model with general error variance–covariance matrix:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{R}\right), \tag{13.20}$$

where $\mathbf{Y}$ is a vector-valued random variable of dimension $r$, $\boldsymbol{\beta}$ is an unknown regression coefficient of dimension $d$, $\mathbf{X}$ is a known $(r \times d)$ design matrix, and $\mathbf{R}$ is a known variance–covariance matrix. In this context filtering refers to inference on $\boldsymbol{\beta}$ through combining of the information contained in a single observation $\mathbf{y}$ from model (13.20) with prior information on $\boldsymbol{\beta}$ expressed through a prior distribution $p(\boldsymbol{\beta})$. Bayes' theorem provides a coherent way of combining these two sources of information by deriving the posterior distribution $p(\boldsymbol{\beta}|\mathbf{R}, \mathbf{y})$:

$$p(\boldsymbol{\beta}|\mathbf{R}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{R}) p(\boldsymbol{\beta}), \tag{13.21}$$

where the likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{R})$ is equal to:

$$p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{R}) = (2\pi)^{-r/2} |\mathbf{R}|^{-1/2} \exp\left( -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{'} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right).$$

For a known variance–covariance matrix $\mathbf{R}$, the likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{R})$ is a quadratic form in $\boldsymbol{\beta}$, hence the conjugate prior $p(\boldsymbol{\beta})$ for the regression coefficient $\boldsymbol{\beta}$ is a normal distribution, $\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$, as is the resulting posterior distribution:

$$\boldsymbol{\beta}|\mathbf{R}, \mathbf{y} \sim \mathcal{N}_d(\mathbf{b}_1, \mathbf{B}_1). \tag{13.22}$$

If $\mathbf{R}^{-1}$ and $\mathbf{B}_0^{-1}$ exist, then the moments of the posterior density are given in terms of the following information filter,

$$\begin{aligned} \mathbf{b}_1 &= \mathbf{B}_1(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{y}), \qquad (13.23)\\ \mathbf{B}_1 &= (\mathbf{B}_0^{-1} + \mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X})^{-1}. \end{aligned}$$

The information filter expresses the posterior mean $\mathbf{b}_1$ as a weighted average of the prior mean $\mathbf{b}_0$ and an estimator that is based entirely on the observation $\mathbf{y}$, with the weights depending on the information obtained in the prior distribution and the likelihood function. If $\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X}$ is invertible, the data-based estimator is equal to the weighted least square estimator $(\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{y}$, and the weight matrices are equal to $\mathbf{B}_1\mathbf{B}_0^{-1}$ and $\mathbf{B}_1\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X}$, respectively.

The information filter involves the inversion of a $(d \times d)$ matrix to obtain the posterior variance–covariance matrix $\mathbf{B}_1$. If the dimension of $\boldsymbol{\beta}$ is larger than the dimension of the observation $\mathbf{y}$ (i.e., $r < d$), or if $\mathbf{R}$ or $\mathbf{B}_0$ are not invertible, it is preferable to work with the following prediction-correction filter which involves the inversion of an $(r \times r)$ matrix, only,

$$\begin{aligned} \mathbf{b}_1 &= \mathbf{b}_0 + \mathbf{K}_1(\mathbf{y} - \mathbf{X}\mathbf{b}_0), \qquad (13.24)\\ \mathbf{B}_1 &= (\mathbf{I}_d - \mathbf{K}_1\mathbf{X})\mathbf{B}_0,\\ \mathbf{K}_1 &= \mathbf{B}_0\mathbf{X}^{'}\mathbf{C}^{-1},\\ \mathbf{C} &= \mathbf{X}\mathbf{B}_0\mathbf{X}^{'} + \mathbf{R}. \qquad (13.25) \end{aligned}$$

The prediction-correction filter expresses the posterior mean $\mathbf{b}_1$ as a correction of the prior mean $\mathbf{b}_0$, which is based on the prediction error $\mathbf{y}-\mathbf{X}\mathbf{b}_0$, resulting from using the prior mean $\mathbf{b}_0$ as an estimator of $\boldsymbol{\beta}$.

It is useful to have an explicit form of the marginal likelihood $p(\mathbf{y}|\mathbf{R})$, that is equal to the normalizing constant of the nonnormalized posterior $p(\boldsymbol{\beta}|\mathbf{R},\mathbf{y})$, given by (13.21):

$$p(\mathbf{y}|\mathbf{R}) = \int p(\mathbf{y}|\boldsymbol{\beta},\mathbf{R})p(\boldsymbol{\beta})d\boldsymbol{\beta}.$$

The marginal likelihood $p(\mathbf{y}|\mathbf{R})$ is obtained from evaluating the following ratio for an arbitrary value of $\boldsymbol{\beta}$,

$$p(\mathbf{y}|\mathbf{R}) = \frac{p(\mathbf{y}|\boldsymbol{\beta},\mathbf{R})p(\boldsymbol{\beta})}{p(\boldsymbol{\beta}|\mathbf{R},\mathbf{y})}.$$

Choosing $\boldsymbol{\beta} = \mathbf{b}_0$ yields

$$p(\mathbf{y}|\mathbf{R}) = (2\pi)^{-r/2}|\mathbf{C}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{b}_0)^{'}\mathbf{C}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}_0)\right), \tag{13.26}$$

which is the density of a multivariate normal distribution with mean $\mathbf{X}\mathbf{b}_0$ and variance–covariance matrix $\mathbf{C}$, when regarded as a function of $\mathbf{y}$.

### 13.3.3 Filtering for the Linear Gaussian State Space Model

For the linear Gaussian state space model defined in (13.6) and (13.7) the posterior density $p(\mathbf{x}_t|\mathbf{y}^t)$ is a normal distribution, where the first two moments are given by the Kalman filter recursions, derived for the first time in Kalman (1960) and Kalman (1961).

*Algorithm 13.1: Kalman Filter* Assume that the filter density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ is the density of a normal distribution:

$$\mathbf{x}_{t-1}|\mathbf{y}^{t-1} \sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}\right). \tag{13.27}$$

Then for a linear Gaussian state space model, the filter density $p(\mathbf{x}_t|\mathbf{y}^t)$ at time $t$ is again the density of a normal distribution obtained from $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ and $\mathbf{y}_t$ through the following steps.

(a) Propagation — determine the density $p(\mathbf{x}_t|\mathbf{y}^{t-1})$:

$$\begin{aligned}
\mathbf{x}_t|\mathbf{y}^{t-1} &\sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{t|t-1}, \mathbf{P}_{t|t-1}\right), \qquad (13.28)\\
\hat{\mathbf{x}}_{t|t-1} &= \mathbf{F}_t\hat{\mathbf{x}}_{t-1|t-1} + \mathbf{G}_t\mathbf{u}_t,\\
\mathbf{P}_{t|t-1} &= \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}_t^{'} + \mathbf{Q}_t.
\end{aligned}$$

(b) Prediction — determine the predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1})$:

$$\begin{aligned} \mathbf{y}_t|\mathbf{y}^{t-1} &\sim \mathcal{N}_r\left(\hat{\mathbf{y}}_{t|t-1}, \mathbf{C}_{t|t-1}\right), \\ \hat{\mathbf{y}}_{t|t-1} &= \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} + \mathbf{A}_t\mathbf{z}_t, \\ \mathbf{C}_{t|t-1} &= \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^{'} + \mathbf{R}_t. \end{aligned} \tag{13.29}$$

(c) Correction — determine the filter density $p(\mathbf{x}_t|\mathbf{y}^t)$:

$$\begin{aligned} \mathbf{x}_t|\mathbf{y}^t &\sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{t|t}, \mathbf{P}_{t|t}\right), \\ \hat{\mathbf{x}}_{t|t} &= \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}), \\ \mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{H}_t^{'}\mathbf{C}_{t|t-1}^{-1}, \\ \mathbf{P}_{t|t} &= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t|t-1}. \end{aligned} \tag{13.30}$$

To start the Kalman filter, one has to choose the normal prior $\mathcal{N}_d\left(\hat{\mathbf{x}}_{0|0}, \mathbf{P}_{0|0}\right)$. It is often recommended to start with a diffuse prior with $\mathbf{P}_{0|0} = \kappa\mathbf{I}_d$ with $\kappa$ being a large value. For state vectors containing both nonstationary and stationary components, De Jong and Chu-Chun-Lin (1994) suggest combining a vague prior with a stationary prior. On the whole, the correct initialization of the Kalman filter is a very subtle issue, and we refer to Koopman (1997) and Durbin and Koopman (2001, Chapter 5) for a very concise and excellent discussion of this issue.

### Derivation of the Kalman Filter

The Kalman filter is easily derived using filtering for a general linear model as in Subsection 13.3.2, as exemplified in Harrison and Stevens (1976) and Meinhold and Singpurwalla (1983).

The density $p(\mathbf{x}_t|\mathbf{y}^{t-1})$ appearing in the propagation step is the normalizing constant of the posterior density $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}^{t-1})$, given by Bayes' theorem as

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}^{t-1}) \propto p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}). \tag{13.31}$$

In (13.31), the transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the likelihood of a general linear model with error variance-covariance matrix $\mathbf{Q}_t$, where the unknown regression parameter $\mathbf{x}_{t-1}$ follows the prior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$, being equal to the filtering density (13.27). The marginal likelihood for this problem is given by (13.26) and takes the form of a normal density in $\mathbf{x}_t$ with the moments being given exactly as in (13.28).

The predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1})$ is the normalizing constant of the filter density $p(\mathbf{x}_t|\mathbf{y}^t)$ which is given by Bayes' theorem:

$$p(\mathbf{x}_t|\mathbf{y}^t) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^{t-1}). \tag{13.32}$$

In (13.32), the observation density $p(\mathbf{y}_t|\mathbf{x}_t)$ is the likelihood of a general linear model with error variance–covariance matrix $\mathbf{R}_t$, where the unknown regression parameter $\mathbf{x}_t$ follows the prior $p(\mathbf{x}_t|\mathbf{y}^{t-1})$ being equal to the propagated density (13.28). Again from Subsection 13.3.2, the posterior $p(\mathbf{x}_t|\mathbf{y}^t)$ is normal with the moments given by (13.30), whereas the marginal likelihood $p(\mathbf{y}_t|\mathbf{y}^{t-1})$ takes the form of a normal density in $\mathbf{y}_t$ with the moments given exactly by (13.29).

For alternative derivations of the Kalman filter based on the concept of projection and minimum mean-squared estimation, see Jazwinski (1970), Anderson and Moore (1979), and Harvey (1989).

### 13.3.4 Filtering for Multiprocess Models

In his pioneering work, Magill (1965) used Bayesian methods to show that for a multiprocess model an explicit solution for the filtering problem is available. If the hidden model indicator $S$ takes $K$ values, then the filter density is a mixture of $K$ normal distributions:

$$p(\mathbf{x}_t|\mathbf{y}^t) = \sum_{k=1}^{K} f_N(\mathbf{x}_t; \hat{\mathbf{x}}_{t|t}^{[k]}, \mathbf{P}_{t|t}^{[k]})\Pr(S = k|\mathbf{y}^t), \qquad (13.33)$$

where the number of components remains fixed for all $t = 1, \ldots, T$. The moments of the various components are obtained by running $K$ parallel Kalman filters as in *Algorithm 13.1*, each conditional on assuming that the state of $S$ is equal to $k$, for $k = 1, \ldots, K$. The component weights are dynamically changing over time and Sims and Lainiotis (1969) showed how they may be updated recursively using Bayes' theorem:

$$\Pr(S = k|\mathbf{y}^t) \propto f_N(\mathbf{y}_t; \hat{\mathbf{y}}_{t|t-1}^{[k]}, \mathbf{C}_{t|t-1}^{[k]})\Pr(S = k|\mathbf{y}^{t-1}),$$

where the moments of the predictive density $p(\mathbf{y}_t|S = k, \mathbf{y}^{t-1})$ are obtained from the Kalman filter corresponding to $S = k$.

### 13.3.5 Approximate Filtering for Switching Linear Gaussian State Space Models

For a switching linear Gaussian state space model the filter density is a mixture of normal distributions:

$$p(\mathbf{x}_t|\mathbf{y}^t) = \qquad (13.34)$$
$$\sum_{(k_1,\ldots,k_t)\in\mathcal{S}_t} f_N(\mathbf{x}_t; \hat{\mathbf{x}}_{t|t}^{[k_1,\ldots,k_t]}, \mathbf{P}_{t|t}^{[k_1,\ldots,k_t]})\Pr(\mathbf{S}^t = (k_1, \ldots, k_t)|\mathbf{y}^t),$$

where $\mathcal{S}_t = \{1, \ldots, K\}^t$ is the space of all paths $\mathbf{S}^t = (S_1, \ldots, S_t)$ up to $t$. This representation holds both for finite mixture as well as Markov switching state space models. In contrast to the multiprocess model, the number of components in the filtering density is increasing exponentially fast. Running an exact recursive filter requires combining all $K^{t-1}$ normal posterior densities $f_N(\mathbf{x}_{t-1}; \hat{\mathbf{x}}_{t-1|t-1}^{[k_1,\dots,k_{t-1}]}, \mathbf{P}_{t-1|t-1}^{[k_1,\dots,k_{t-1}]})$ with each of the $K$ states of $S_t$, running in total $K^t$ parallel Kalman filters as in *Algorithm 13.1.* This is operational only if the total number $T$ of observations is not too large; see, for instance, Schervish and Tsay (1988) for an empirical application of this filter.

In most cases some approximate filter has to be applied. Approximate filters for switching Gaussian state space models were studied rather early in the engineering literature; we mention here in particular Ackerson and Fu (1970), Bar-Shalom and Tse (1975), Akashi and Kumamoto (1977), Tugnait (1982), and Blom and Bar-Shalom (1988). Approximations in the statistical and econometric literature were suggested by Harrison and Stevens (1976), Cosslett and Lee (1985), Peña and Guttman (1988), Lam (1990), Gordon and Smith (1990), Shumway and Stoffer (1991), and Kim (1994). To keep the filter operational, the number of components of the filtering density has to be limited, usually by merging components at each filter step. Other techniques are trimming by removing unlikely components with small probability and combining similar components into a single component.

A useful starting point for discussing the various approximate filters is writing the filter density $p(\mathbf{x}_t|\mathbf{y}^t)$ as

$$p(\mathbf{x}_t|\mathbf{y}^t) = \sum_{k=1}^{K} p(\mathbf{x}_t|\mathbf{y}^t, S_t = k)\mathrm{Pr}(S_t = k|\mathbf{y}^t). \tag{13.35}$$

In (13.35) we identify two filtering problems. First, we need to derive the discrete filter probabilities $\mathrm{Pr}(S_t = k|\mathbf{y}^t)$ for $k = 1, \dots, K$ without conditioning on the continuous state vector $\mathbf{x}_t$; second, we need to derive filter recursion for the continuous state $\mathbf{x}_t$ conditional on knowing only the present state of $S_t$.

For a hidden Markov chain $S_t$ with transition matrix $\boldsymbol{\xi}$, the discrete filter is derived through Bayes' theorem in a similar way as was done for Markov switching models in Section 11.2:

$$\mathrm{Pr}(S_t = k|\mathbf{y}^t) \propto p(\mathbf{y}_t|S_t = k, \mathbf{y}^{t-1})\mathrm{Pr}(S_t = k|\mathbf{y}^{t-1}). \tag{13.36}$$

The propagated probabilities $\mathrm{Pr}(S_t = k|\mathbf{y}^{t-1})$ are essentially the same as in Section 11.2 and read:

$$\mathrm{Pr}(S_t = k|\mathbf{y}^{t-1}) = \sum_{j=1}^{K} \xi_{jk}\mathrm{Pr}(S_{t-1} = j|\mathbf{y}^{t-1}).$$

For a hidden i.i.d. indicator this reduces to $\mathrm{Pr}(S_t = k|\mathbf{y}^{t-1}) = \eta_k$ as $\xi_{jk} = \eta_k$. Because the likelihood $p(\mathbf{y}_t|S_t = k, \mathbf{y}^{t-1})$ in (13.36) will also appear in the prediction step of the second filtering problem, both filtering problems are related.

To solve the second filtering problem, a recursion between the filter densities $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j)$ and $p(\mathbf{x}_t|\mathbf{y}^t, S_t = k)$ has to be established. One could, in principle, proceed as in Subsection 13.3.3, using the propagation step

$$\begin{aligned} &p(\mathbf{x}_t|\mathbf{y}^{t-1}, S_t = k) = \\ &\int p(\mathbf{x}_t|\mathbf{x}_{t-1}, S_t = k)p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)d\mathbf{x}_{t-1}, \end{aligned} \tag{13.37}$$

the prediction step

$$p(\mathbf{y}_t|S_t = k, \mathbf{y}^{t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t, S_t = k)p(\mathbf{x}_t|\mathbf{y}^{t-1}, S_t = k)d\mathbf{x}_t, \tag{13.38}$$

and the correction step

$$p(\mathbf{x}_t|\mathbf{y}^t, S_t = k) \propto p(\mathbf{y}_t|\mathbf{x}_t, S_t = k)p(\mathbf{x}_t|\mathbf{y}^{t-1}, S_t = k). \tag{13.39}$$

Because we are dealing with a finite or Markov mixture of linear Gaussian state space models, the transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1}, S_t = k)$ and the observation density $p(\mathbf{y}_t|\mathbf{x}_t, S_t = k)$ are normal, however, $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ does not have the required form of a conjugate normal prior. Nonnormality of $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ arises due to possible changes in the states of $S_{t-1}$ and $S_t$ between $t-1$ and $t$, which may occur both for finite mixture as well as Markov switching state space models. $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ is a finite mixture of the filtering densities at $t-1$:

$$p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k) = \sum_{j=1}^{K} p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j)w_{jk}, \tag{13.40}$$

where the weights are given by

$$w_{jk} = \Pr(S_{t-1} = j|\mathbf{y}^{t-1}, S_t = k) \propto \xi_{jk}\Pr(S_{t-1} = j|\mathbf{y}^{t-1}).$$

For a Markov switching state space model, the weights read:

$$w_{jk} = \frac{\xi_{jk}\Pr(S_{t-1} = j|\mathbf{y}^{t-1})}{\sum_{l=1}^{K} \xi_{lk}\Pr(S_{t-1} = l|\mathbf{y}^{t-1})}. \tag{13.41}$$

For a finite mixture state space model, the weights are identical with the discrete filter probabilities:

$$w_{jk} = \Pr(S_{t-1} = j|\mathbf{y}^{t-1}). \tag{13.42}$$

In principle, these formulae provide a recursion comparable to the Kalman filter. However, because the filter density $p(\mathbf{x}_t|\mathbf{y}^t, S_t = k)$ is given by a mixture of $K_t = KK_{t-1}$ components, where $K_{t-1}$ is the number of components at $t-1$, some method of limiting the number of components must be found to make this filter operational. As pointed out by Blom and Bar-Shalom (1988), different algorithms emerge, depending on the precise density and the precise time point chosen for this simplification.

### Kim's Algorithm

This algorithm was suggested independently by Tugnait (1982) and Kim (1994). Assume that the filter density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j)$ is a normal distribution:

$$p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j) = f_N(\mathbf{x}_{t-1}; \hat{\mathbf{x}}^{[j]}_{t-1|t-1}, \mathbf{P}^{[j]}_{t-1|t-1}). \qquad (13.43)$$

Then the prior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ in (13.40) is a mixture of $K$ normal distributions as is the filter density $p(\mathbf{x}_t|\mathbf{y}^t, S_t = k)$ in (13.39):

$$\begin{aligned} &p(\mathbf{x}_t|\mathbf{y}^t, S_t = k) = \qquad (13.44)\\ &\sum_{j=1}^{K} f_N(\mathbf{x}_t; \hat{\mathbf{x}}^{[j,k]}_{t|t}, \mathbf{P}^{[j,k]}_{t|t}) \Pr(S_{t-1} = j|\mathbf{y}^t, S_t = k). \end{aligned}$$

The component densities in the filter density are obtained by running in total $K^2$ Kalman filters, combining each normal density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j)$ with each possible value for $S_t = k$. Each Kalman filter delivers the normal one-step ahead predictive density

$$p(\mathbf{y}_t|\mathbf{y}^{t-1}, S_{t-1} = j, S_t = k) = f_N(\mathbf{y}_t; \hat{\mathbf{y}}^{[j,k]}_{t|t-1}, \mathbf{C}^{[j,k]}_{t|t-1}),$$

which could be used to compute the weights $\Pr(S_{t-1} = j|\mathbf{y}^t, S_t = k)$ in (13.44) through Bayes' theorem:

$$\begin{aligned} &\Pr(S_{t-1} = j|\mathbf{y}^t, S_t = k) \propto \qquad (13.45)\\ &\quad p(\mathbf{y}_t|\mathbf{y}^{t-1}, S_{t-1} = j, S_t = k) w_{jk}, \end{aligned}$$

where $w_{jk}$ were defined in (13.41) and (13.42), respectively. For each value of $k$, the normalizing constant of the right-hand side of (13.45) is equal to the one-step ahead predictive density $p(\mathbf{y}_t|S_t = k, \mathbf{y}^{t-1})$,

$$p(\mathbf{y}_t|S_t = k, \mathbf{y}^{t-1}) = \sum_{j=1}^{K} p(\mathbf{y}_t|\mathbf{y}^{t-1}, S_{t-1} = j, S_t = k) w_{jk},$$

which is necessary for the computation of the discrete filter probabilities $\Pr(S_t = k|\mathbf{y}^t)$ through (13.36).

To keep the filter operational, Kim (1994) collapses the mixture (13.44) to a single normal density after having finished filtering at time $t$, which it is then used as a prior density for the next filtering step:

$$p(\mathbf{x}_t|\mathbf{y}^t, S_t = k) \approx f_N(\mathbf{x}_t; \hat{\mathbf{x}}^{[k]}_{t|t}, \mathbf{P}^{[k]}_{t|t}),$$

$$\hat{\mathbf{x}}^{[k]}_{t|t} = \sum_{j=1}^{K} \hat{\mathbf{x}}^{[j,k]}_{t|t} \Pr(S_{t-1} = j|\mathbf{y}^t, S_t = k),$$

$$\mathbf{P}^{[k]}_{t|t} = \sum_{j=1}^{K} (\hat{\mathbf{x}}^{[j,k]}_{t|t} (\hat{\mathbf{x}}^{[j,k]}_{t|t})' + \mathbf{P}^{[j,k]}_{t|t}) \Pr(S_{t-1} = j|\mathbf{y}^t, S_t = k) - \hat{\mathbf{x}}^{[k]}_{t|t} (\hat{\mathbf{x}}^{[k]}_{t|t})'.$$

A comparison of this approximate filter with exact inference in Kim (1994) for the model of Lam (1990) indicates that this approximate filter is quite accurate.

Tugnait (1982) extended this method by updating a whole sequence $(S_{t-h}, \dots, S_t)$ with $h > 1$.

### Other Approximations

Several other approximations also assume that the prior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j)$ is a normal distribution as in (13.43), reduction of filter complexity, however, is carried out in a different manner. Blom and Bar-Shalom (1988) suggest collapsing the mixture density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ given by (13.40) to a single normal density with the same moments *prior* to running through the filter steps (13.37) to (13.38) at time $t$:

$$\begin{aligned} p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k) &\approx f_N(\mathbf{x}_t; \hat{\mathbf{x}}^{[k]}_{t-1|t-1}, \mathbf{P}^{[k]}_{t-1|t-1}) \\ \hat{\mathbf{x}}^{[k]}_{t-1|t-1} &= \sum_{j=1}^{K} w_{jk} \mathrm{E}(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_{t-1} = j), \end{aligned} \tag{13.46}$$

with a similar formula for the variance–covariance matrix. Filtering then reduces to running $K$ Kalman filters, however, this filter is less precise than Kim's algorithm.

For finite mixture of state space models, the weights in (13.46) are independent of $k$, $w_{jk} = \Pr(S_{t-1} = j|\mathbf{y}^{t-1})$ (see again (13.42)), and all moments in (13.46) reduce to the moments $\hat{\mathbf{x}}_{t-1|t-1}$ and $\mathbf{P}_{t-1|t-1}$ of the marginal posterior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$,

$$p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k) \approx f_N(\mathbf{x}_t; \hat{\mathbf{x}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}).$$

Such a filter is running through the filter steps (13.37) to (13.38) with the same prior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, S_t = k)$ for all $k$ and reduces the collapsing procedures suggested by Harrison and Stevens (1976), Peña and Guttman (1988), and Shumway and Stoffer (1991) for finite mixtures of state space models.

Ackerson and Fu (1970) and Bar-Shalom and Tse (1975) use the same collapsing technique, where the unconditional posterior $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ is approximated by a single normal density prior to filtering also for Markov switching state space models. This procedure, however, is likely to be less optimal than the collapsing method of Blom and Bar-Shalom (1988), especially for highly persistent Markov chains, whereas there is little computational gain.

## 13.4 Parameter Estimation for Switching State Space Models

Let $\boldsymbol{\vartheta}$ summarize all unknown distinct parameters appearing in the definition of a switching state space model that should be fitted to a univariate or multivariate time series $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)$. In various applications of switching state space models, the parameters of the probability law of $S_t$ and the covariances $\mathbf{Q}_t$ and $\mathbf{R}_t$ are assumed to be known, often based by choosing somewhat arbitrary values (Harrison and Stevens, 1976; Carter and Kohn, 1994), but in general these parameters may be estimated from the data as well.

### 13.4.1 The Likelihood Function of a State Space Model

The likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$ is defined as the density $p(\mathbf{y}_1, \dots, \mathbf{y}_T|\boldsymbol{\vartheta})$ of the joint distribution of $\mathbf{Y}_1, \dots, \mathbf{Y}_T$ where all latent variables, in particular the state process $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_T)$ and the indicator process $\mathbf{S} = (S_0, \dots, S_T)$, are integrated out. In general, the likelihood of a state space model is derived by using the following decomposition into one-step ahead predictive densities (Schweppe, 1965; Kashyap, 1970),

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = p(\mathbf{y}_1, \dots, \mathbf{y}_T|\boldsymbol{\vartheta}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}).$$

For a linear Gaussian state space model the predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ appears as part of the Kalman filter (see *Algorithm 13.1*), and the likelihood function is obtained from a single run of the Kalman filter conditional on $\boldsymbol{\vartheta}$, if the initial moments $\hat{\mathbf{x}}_{0|0}$ and $\mathbf{P}_{0|0}$ are known:

$$\begin{aligned}&-2\log p(\mathbf{y}_1, \dots, \mathbf{y}_T|\boldsymbol{\vartheta})\\ &= \sum_{t=1}^{T} \left( \log |\mathbf{C}_{t|t-1}(\boldsymbol{\vartheta})| + (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\vartheta}))' \mathbf{C}_{t|t-1}(\boldsymbol{\vartheta})^{-1} (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\vartheta})) \right),\end{aligned}$$

where $\hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\vartheta})$ and $\mathbf{C}_{t|t-1}(\boldsymbol{\vartheta})$ are given by (13.29). Some care needs to be exercised if the initial moments $\hat{\mathbf{x}}_{0|0}$ and $\mathbf{P}_{0|0}$ are unknown, and we refer to Durbin and Koopman (2001, Section 7.2) for further discussion.

For a switching linear Gaussian state space model, the likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ where both sets of latent variables are integrated out is not available in closed form. Like the filter density $p(\mathbf{x}_t|\mathbf{y}^t, \boldsymbol{\vartheta})$, the one-step ahead predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ is a mixture of normal densities with an increasing number of components. However, any of the approximate filters discussed in Subsection 13.3.5 leads immediately to an approximation to the log likelihood function. By rewriting the predictive density as

$$p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(\mathbf{y}_t|\mathbf{y}^{t-1}, S_t = k, \boldsymbol{\vartheta}) \Pr(S_t = k|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}),$$

it becomes evident that $p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ is the normalizing constant of the right-hand side of discrete filter distribution $\Pr(S_t = k|\mathbf{y}, \boldsymbol{\vartheta})$, given in (13.36). Approximate ML estimation based on approximate filters has been applied by Shumway and Stoffer (1991) and Kim (1994), among others.

It is worth noting that certain partial likelihood functions are available in closed form. When holding $\mathbf{S}$ fixed, one is dealing with a standard state space model, and the likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}, \mathbf{S})$ is obtained by running a Kalman filter conditional on $\boldsymbol{\vartheta}$ and $\mathbf{S}$.

### 13.4.2 Maximum Likelihood Estimation

A straightforward method of obtaining the ML estimator is direct maximization of the exact or approximate log likelihood function $\log p(\mathbf{y}_1, \dots, \mathbf{y}_T|\boldsymbol{\vartheta})$ using some numerical technique such as Newton–Raphson methods; see, for instance, Hamilton (1994b, Section 5.7) for a review of these methods.

It was realized by Shumway and Stoffer (1982) and Watson and Engle (1983) that the EM algorithm of Dempster et al. (1977) may be applied to linear Gaussian state space models without switching, because the complete-data likelihood function $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\vartheta})p(\mathbf{x}|\boldsymbol{\vartheta})$ turns out to be of simple form. Koopman (1993) proposed a very simple and efficient EM algorithm for unknown parameters inside the variance–covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$ of a linear Gaussian state space form.

For a switching state space model, the presence of two sets of latent variables hinders a straightforward application of the EM algorithm, because the required smoothed probabilities $\Pr(S_t = k|\mathbf{y})$ are not available in closed form. Shumway and Stoffer (1991) substitute these probabilities by $\Pr(S_t = k|\mathbf{y}^t)$ which are available from any approximate filter discussed in Subsection 13.3.5 and report that this pseudo EM algorithm works well.

Consistency and asymptotic normality of the ML estimator of the parameters of a state space model hold under fairly general conditions; see Shumway and Stoffer (1982), Schneider (1988), Hamilton (1994b, Section 13.4), Jensen and Petersen (1999), and Shumway and Stoffer (2000, p.326ff). The observed time series, however, needs to be fairly long in order to achieve asymptotic normality. Moreover, problems occur if some of the parameters are close to the boundary of the parameter space. For this reason it seems sensible to consider a Bayesian approach.

### 13.4.3 Bayesian Inference

Bayesian inference for switching state space models is based on deriving the joint posterior density $p(\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y})$ of all continuous states $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_T)$, all discrete states $\mathbf{S} = (S_0, \dots, S_T)$, and unknown model parameters $\boldsymbol{\vartheta}$, including unknown parameters in the probability law of $\mathbf{S}$, if any are present. Due to the hierarchical structure of a switching state space model, this density is proportional to:

$$p(\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{x}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}),$$

which simplifies to:

$$p(\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y}) \propto p(\mathbf{x}_0|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}) \times \prod_{t=1}^{N} p(\mathbf{y}_t|S_t, \mathbf{x}_t, \boldsymbol{\vartheta})p(\mathbf{x}_t|S_t, \mathbf{x}_{t-1}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\vartheta}). \tag{13.47}$$

The densities $p(\mathbf{y}_t|S_t, \mathbf{x}_t, \boldsymbol{\vartheta})$ and $p(\mathbf{x}_t|S_t, \mathbf{x}_{t-1}, \boldsymbol{\vartheta})$ result directly from the definition of the state space model, where $p(\mathbf{x}_0|\boldsymbol{\vartheta})$ is the prior of $\mathbf{x}_0$. $p(\boldsymbol{\vartheta})$ is the prior density of all model parameters. The density $p(\mathbf{S}|\boldsymbol{\vartheta})$ results directly from the definition of the probability law of $S_t$. If $S_t$ is a hidden Markov chain, then

$$p(\mathbf{S}|\boldsymbol{\vartheta}) = p(S_0|\boldsymbol{\vartheta}) \prod_{t=1}^{N} p(S_t|S_{t-1}, \boldsymbol{\vartheta}).$$

If $S_t$ is a hidden i.i.d. indicator, then

$$p(\mathbf{S}|\boldsymbol{\vartheta}) = \prod_{t=1}^{N} p(S_t|\boldsymbol{\vartheta}).$$

Note that the derivation of the posterior density in (13.47) is not limited to switching linear Gaussian state space models, but is valid for any switching state space model.

The posterior density $p(\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta}|\mathbf{y})$, however, is not of any closed form, even for linear Gaussian state space models without switching and simulation-based methods are usually applied for Bayesian estimation. Durbin and Koopman (2000) propagate the application of importance sampling, several other authors explored MCMC methods; see Section 13.5.

### Choosing the Priors for Bayesian Estimation

If $S_t$ is a hidden Markov chain with transition matrix $\boldsymbol{\xi}$, then the joint prior reads

$$p(\mathbf{x}_0|\boldsymbol{\vartheta}, S_0)p(S_0|\boldsymbol{\xi})p(\boldsymbol{\vartheta})p(\boldsymbol{\xi}), \tag{13.48}$$

where each row $\boldsymbol{\xi}_{j\cdot}$ of the transition matrix $\boldsymbol{\xi}$ is chosen from a Dirichlet distribution as in Chapter 11:

$$\boldsymbol{\xi}_{k\cdot} \sim \mathcal{D}\left(e_{k1}, \ldots, e_{kK}\right), \qquad k = 1, \ldots, K. \tag{13.49}$$

To obtain a prior that is invariant to relabeling, Frühwirth-Schnatter (2001a) suggested choosing $e_{kk} = e^P$ and $e_{kk'} = e^T$, if $k \neq k'$. By choosing $e^P > e^T$, a Markov switching state space model is bounded away from a finite mixture state space model. Choosing the prior $p(S_0|\boldsymbol{\xi})$ of the discrete-valued state variable $S_0$ is closely related to choosing the same prior for finite Markov mixture models; see Subsection 10.3.4 for various choices of this distribution.

If $S_t$ is a hidden i.i.d. indicator with probability distribution $\boldsymbol{\eta}$, then the joint prior reduces to

$$p(\mathbf{x}_0|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})p(\boldsymbol{\eta}), \tag{13.50}$$

where the prior for $\boldsymbol{\eta}$ is chosen from the Dirichlet distribution as in Chapter 2:

$$\boldsymbol{\eta} \sim \mathcal{D}\left(e_0, \dots, e_0\right). \tag{13.51}$$

In both cases, $p(\mathbf{x}_0|\boldsymbol{\vartheta}, S_0)$ is the prior for the continuous state variable $\mathbf{x}_0$ used for initialization in the Kalman filter, and is allowed to depend on $S_0$ for a Markov switching state space model.

The prior for the remaining parameters $\boldsymbol{\vartheta}$ is usually chosen to be conditionally conjugate to the complete-data likelihood $p(\mathbf{y}|\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{x}|\mathbf{S}, \boldsymbol{\vartheta})$. To give an example, consider a local level model where both variances are switching,

$$\begin{aligned} \mu_t &= \mu_{t-1} + w_t, & w_t &\sim \mathcal{N}\left(0, \sigma^2_{\mu,S_t}\right), \\ Y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_t}\right). \end{aligned} \tag{13.52}$$

The complete-data likelihood reads with $\mathbf{x} = (\mu_0, \dots, \mu_T)$ and $\boldsymbol{\vartheta} = (\sigma^2_{\mu,1}, \dots, \sigma^2_{\mu,K}, \sigma^2_{\varepsilon,1}, \dots, \sigma^2_{\varepsilon,K})$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{x}|\mathbf{S}, \boldsymbol{\vartheta}) \propto \prod_{k=1}^{K} \left(\frac{1}{\sigma^2_{\varepsilon,k}}\right)^{N_k(\mathbf{S})/2} &\exp\left\{-\frac{\sum_{t:S_t=k} (y_t - \mu_t)^2}{2\sigma^2_{\varepsilon,k}}\right\} \\ \times \left(\frac{1}{\sigma^2_{\mu,k}}\right)^{N_k(\mathbf{S})/2} &\exp\left\{-\frac{\sum_{t:S_t=k} (\mu_t - \mu_{t-1})^2}{2\sigma^2_{\mu,k}}\right\}, \end{aligned}$$

where $N_k(\mathbf{S}) = \#\{S_t = k\}$. Considered as a function of $\sigma^2_{\varepsilon,k}$, this is an inverted Gamma density. Therefore the conditionally conjugate prior for $\sigma^2_{\varepsilon,k}$ is an inverted Gamma density $\mathcal{G}^{-1}\left(c_{\varepsilon,0}, C_{\varepsilon,0}\right)$. Similarly, the complete-data likelihood is an inverted Gamma density, when considered as a function of $\sigma^2_{\mu,k}$. Thus the conditionally conjugate prior for $\sigma^2_{\mu,k}$ is again an inverted Gamma density $\mathcal{G}^{-1}\left(c_{\mu,0}, C_{\mu,0}\right)$.

### Complete-Data Bayesian Estimation

Estimation of the unknown model parameters $\boldsymbol{\vartheta}$ conditional on the complete data $\mathbf{S}$, $\mathbf{x}$, and $\mathbf{y}$ is closely related to various Bayesian inference problems discussed earlier. If parameters appearing in the definition of the probability law of $S_t$ are a priori independent of parameters appearing in the definition of the transition and observation densities, then this independence is preserved a posteriori. If $S_t$ is an i.i.d. indicator with unknown probability distribution $\boldsymbol{\eta}$, then $\boldsymbol{\eta}|\mathbf{S}, \mathbf{x}, \mathbf{y}$ follows a Dirichlet distribution as discussed for finite mixture models in Subsection 3.5.3, whereas the posterior of $\boldsymbol{\xi}|\mathbf{S}, \mathbf{x}, \mathbf{y}$ under a hidden Markov chain $S_t$ with unknown transition matrix $\boldsymbol{\xi}$ is the same as in Subsection 11.5.5. For unknown parameters appearing in the definition of the observation and the transition density, the complete-data likelihood $p(\mathbf{y}|\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{x}|\mathbf{S}, \boldsymbol{\vartheta})$ in combination with a conditionally conjugate prior $p(\boldsymbol{\vartheta})$ often leads to a posterior density $p(\boldsymbol{\vartheta}|\mathbf{S}, \mathbf{x}, \mathbf{y})$ that is of closed form.

To give an example, consider a local level model where both variances are switching as in (13.52) and $S_t$ is a hidden Markov chain. Then $\boldsymbol{\vartheta} = (\sigma^2_{\mu,1}, \dots, \sigma^2_{\mu,K}, \sigma^2_{\varepsilon,1}, \dots, \sigma^2_{\varepsilon,K}, \boldsymbol{\xi})$ and the complete-data posterior $p(\boldsymbol{\vartheta}|\mathbf{x}, \mathbf{S}, \mathbf{y})$ reads:

$$p(\boldsymbol{\vartheta}|\mathbf{x}, \mathbf{S}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{x}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\xi})p(\boldsymbol{\vartheta}) \propto p(S_0|\boldsymbol{\vartheta}) \prod_{j=1}^{K} \prod_{k=1}^{K} \xi_{jk}^{N_{jk}(\mathbf{S})}$$

$$\times \prod_{k=1}^{K} \left(\frac{1}{\sigma^2_{\varepsilon,k}}\right)^{N_k(\mathbf{S})/2+c_{\varepsilon,0}+1} \exp\left\{-\frac{\sum_{t:S_t=k} (y_t - \mu_t)^2}{2\sigma^2_{\varepsilon,k}} - \frac{C_{\varepsilon,0}}{\sigma^2_{\varepsilon,k}}\right\}$$

$$\times \prod_{k=1}^{K} \left(\frac{1}{\sigma^2_{\mu,k}}\right)^{N_k(\mathbf{S})/2+c_{\mu,0}+1} \exp\left\{-\frac{\sum_{t:S_t=k} (\mu_t - \mu_{t-1})^2}{2\sigma^2_{\mu,k}} - \frac{C_{\mu,0}}{\sigma^2_{\mu,k}}\right\},$$

where $N_{jk}(\mathbf{S}) = \# \{S_{t-1} = j, S_t = k\}$ counts the numbers of transitions from $j$ to $k$ and $N_k(\mathbf{S}) = \#\{S_t = k\} = \sum_{j=1}^{K} N_{jk}(\mathbf{S})$. The transition matrix $\boldsymbol{\xi}$, as well as all variances $\sigma^2_{\mu,k}$ and $\sigma^2_{\varepsilon,k}$ are conditionally independent. The precise form of the posterior of $\boldsymbol{\xi}$ and the method used for sampling from this density depend on the assumptions concerning $p(S_0|\boldsymbol{\vartheta})$, as has been discussed earlier in Subsection 11.5.5. The variances $\sigma^2_{\mu,k}$ and $\sigma^2_{\varepsilon,k}$ each follow an inverted Gamma density $\mathcal{G}^{-1}(c_{\mu,k}(\mathbf{S}), C_{\mu,k}(\mathbf{S}))$ and $\mathcal{G}^{-1}(c_{\varepsilon,k}(\mathbf{S}), C_{\varepsilon,k}(\mathbf{S}))$, where

$$c_{\varepsilon,k}(\mathbf{S}) = c_{\varepsilon,0} + 0.5N_k(\mathbf{S}), \qquad C_{\varepsilon,k}(\mathbf{S}) = C_{\varepsilon,0} + 0.5 \sum_{t:S_t=k} (y_t - \mu_t)^2,$$

$$c_{\mu,k}(\mathbf{S}) = c_{\mu,0} + 0.5N_k(\mathbf{S}), \qquad C_{\mu,k}(\mathbf{S}) = C_{\mu,0} + 0.5 \sum_{t:S_t=k} (\mu_t - \mu_{t-1})^2.$$

## 13.5 Practical Bayesian Estimation Using MCMC

Practical Bayesian estimation of switching state space models usually relies on MCMC estimation and was implemented for specific models discussed in Section 13.2 such as the state space model with Markov switching conditional heteroscedasticity (Carlin et al., 1992; Carter and Kohn, 1994, 1996), the random level shift model (McCulloch and Tsay, 1993), partial Gaussian state space model (Shephard, 1994), robust state space model (Godsill, 1997; Godsill and Rayner, 1998), dynamic factor model with regime switching (Kim and Nelson, 1998; Kaufmann, 2000), and various unobserved component models with Markov switching (Luginbuhl and de Vos, 1999; Engel and Kim, 1999). Frühwirth-Schnatter (2001a) provides a general discussion of MCMC methods for switching linear Gaussian state space models.

### 13.5.1 Various Data Augmentation Schemes

Various MCMC schemes have been suggested to implement data augmentation and Gibbs sampling for switching linear Gaussian state space models. The following three-block Gibbs sampler has been applied in Shephard (1994), Carter and Kohn (1994), and Frühwirth-Schnatter (2001a).

*Algorithm 13.2: MCMC for a Switching Linear Gaussian State Space Model — Full Conditional Gibbs Sampling* Sampling is carried out in three steps.

(a) Sample a path $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_T)$ of the continuous state variable conditional on $\boldsymbol{\vartheta}$ and $\mathbf{S}$ from the density $p(\mathbf{x}|\boldsymbol{\vartheta}, \mathbf{S}, \mathbf{y})$, preferably using forward-filtering-backward-sampling; see *Algorithm 13.4*.
(b) Sample a path $\mathbf{S} = (S_0, \dots, S_T)$ of the discrete state variable conditional on $\boldsymbol{\vartheta}$ and $\mathbf{x}$ from the density $p(\mathbf{S}|\boldsymbol{\vartheta}, \mathbf{x}, \mathbf{y})$.
(c) Sample $\boldsymbol{\vartheta}$ conditional on $\mathbf{x}$ and $\mathbf{S}$ from the complete-data posterior density $p(\boldsymbol{\vartheta}|\mathbf{x}, \mathbf{S}, \mathbf{y})$.

Sampling a path of the state process $\mathbf{x}_0, \dots, \mathbf{x}_T$ in step (a) is discussed in full detail in Subsection 13.5.2. Sampling the indicators in step (b) is straightforward, if $S_t$ is a hidden i.i.d. sequence with probability distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$. In this case, $S_t$ is independent of all other indicators $\mathbf{S}_{-t}$ given $\mathbf{x}$, and step (b) could be carried out in one sweep by sampling $S_t$ for each $t = 1, \dots, T$ from

$$\begin{aligned}&\Pr(S_t = j|\mathbf{y}, \mathbf{x}, \boldsymbol{\vartheta}) \\ &\propto p(\mathbf{y}_t|S_t = j, \mathbf{x}_t, \boldsymbol{\vartheta}) p(\mathbf{x}_t|S_t = j, \mathbf{x}_{t-1}, \boldsymbol{\vartheta}) \eta_k.\end{aligned} \tag{13.53}$$

If $S_t$ is a hidden Markov chain, then the results derived earlier for sampling hidden Markov chains are extended to deal with switching state space models; see *Algorithm 13.5* for more details. Sampling the unknown model parameters in step (c) conditional on $\mathbf{S}$, $\mathbf{x}$, and $\mathbf{y}$ has been discussed earlier in Subsection 13.4.3.

Carter and Kohn (1996, Lemma 2.2) prove that full conditional Gibbs sampling may lead to a reducible sampler for certain state space models. This is the case, for instance, if one of the variances, say $\mathbf{Q}_t^{[k]}$, is assumed to be exactly 0, if $S_t = k$. As a remedy, Carter and Kohn (1996) substitute step

(b) in *Algorithm 13.2* by a step that samples $S_t$ without conditioning on the continuous states $\mathbf{x}$.

*Algorithm 13.3: MCMC for a Switching Linear Gaussian State Space Model — Marginal Sampling of the Indicators* Whereas sampling of $\mathbf{x}$ and $\boldsymbol{\vartheta}$ is the same as in step (a) and (c) in *Algorithm 13.2*, marginal sampling of the indicators is carried out in the following way.

(b) For $t = 1, \ldots, T$, sample $S_t$ from $p(S_t|\mathbf{S}_{-t}, \boldsymbol{\vartheta}, \mathbf{y})$ without conditioning on $\mathbf{x}$.

Generating the indicators $S_t$ in step (b) of this algorithm in an efficient way is far from straightforward. Carter and Kohn (1996) and Gerlach and Kohn (2000) discuss various samplers, that are reviewed in Subsection 13.5.3. The results of Liu et al. (1994) suggest that *Algorithm 13.3* is more efficient than *Algorithm 13.2*, because the indicators are conditioned on fewer variables when they are generated. This is supported by a small simulation study in Gerlach and Kohn (2000).

Another modification of *Algorithm 13.2* is a partially marginalized sampler (McCulloch and Tsay, 1993; Godsill, 1997; Godsill and Rayner, 1998), where sampling of the indicators and the states is carried out in a different manner.

## 13.5.2 Sampling the Continuous State Process from the Smoother Density

In this section, sampling a path of the state process $\mathbf{x}_0, \ldots, \mathbf{x}_T$ from the conditional posterior $p(\mathbf{x}_0, \ldots, \mathbf{x}_T|\mathbf{y}, \mathbf{S}, \boldsymbol{\vartheta})$, also called smoother density, is discussed in full detail. The transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ as well as the observation density $p(\mathbf{y}_t|\mathbf{x}_t)$ depends on unknown parameters $\boldsymbol{\vartheta}$ and the latent processes $\mathbf{S}$. This dependence, however, is dropped for the remainder of this subsection for notational convenience.

### Single-Move Sampling of the Continuous State Process

Carlin et al. (1992) used a single-move Gibbs sampler based on sampling the state $\mathbf{x}_t$ for each $t = 1, \ldots, T$ from the conditional posterior $\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{\_t}, \mathbf{y})$, where $\mathbf{x}_{\_t}$ is the collection all state vectors $\mathbf{x}_0, \ldots, \mathbf{x}_T$ excluding $\mathbf{x}_t$. The posterior $p(\mathbf{x}_t|\mathbf{x}_{\_t}, \mathbf{y})$ is given by

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_{\_t}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \\ &\propto \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_0). \end{aligned}$$

Dropping all quantities that are independent of $\mathbf{x}_t$ yields for $t = 1, \ldots, T-1$:

$$p(\mathbf{x}_t|\mathbf{x}_{\_t}, \mathbf{y}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{13.54}$$

with obvious simplifications for $t = 0$ and $t = T$:

$$\begin{aligned} &p(\mathbf{x}_0|\mathbf{x}_1, \dots, \mathbf{x}_T, \mathbf{y}) \propto p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_0), \\ &p(\mathbf{x}_T|\mathbf{x}_0, \dots, \mathbf{x}_{T-1}, \mathbf{y}) \propto p(\mathbf{y}_T|\mathbf{x}_T)p(\mathbf{x}_T|\mathbf{x}_{T-1}). \end{aligned}$$

For a linear Gaussian state space model the first two densities in (13.54) may be considered as the likelihood of a linear model with general, but known, error covariance matrices and independent observations $\mathbf{y}_t$ and $\mathbf{x}_{t+1}$,

$$\begin{gathered} \begin{pmatrix} \mathbf{y}_t \\ \mathbf{x}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_t \\ \mathbf{F}_{t+1} \end{pmatrix} \mathbf{x}_t + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{w}_{t+1} \end{pmatrix}, \\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}_r\left(\mathbf{0}, \mathbf{R}_t\right), \qquad \mathbf{w}_{t+1} \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}_{t+1}\right), \end{gathered}$$

where the unknown regression parameter $\mathbf{x}_t$ follows the conjugate normal prior $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ as in Subsection 13.3.2. Thus for $t = 1, \dots, T-1$ the density $p(\mathbf{x}_t|\mathbf{x}_{-t}, \mathbf{y})$ is normal with

$$\begin{aligned} &\mathbf{x}_t|\mathbf{x}_{-t}, \mathbf{y} \sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{t|-t}, \mathbf{P}_{t|-t}\right), \\ &\mathbf{P}_{t|-t}^{-1} = \mathbf{H}_t^{'}\mathbf{R}_t^{-1}\mathbf{H}_t + \mathbf{F}_{t+1}^{'}\mathbf{Q}_{t+1}^{-1}\mathbf{F}_{t+1} + \mathbf{Q}_t^{-1}, \\ &\hat{\mathbf{x}}_{t|-t} = \mathbf{P}_{t|-t}(\mathbf{H}_t^{'}\mathbf{R}_t^{-1}\mathbf{y}_t + \mathbf{F}_{t+1}^{'}\mathbf{Q}_{t+1}^{-1}\mathbf{x}_{t+1} + \mathbf{Q}_t^{-1}\mathbf{F}_t\mathbf{x}_{t-1}), \end{aligned}$$

a result that allows direct sampling. For more general state space models, $p(\mathbf{x}_t|\mathbf{x}_{-t}, \mathbf{y})$ is no longer a normal density, but it is possible to draw from this density using a Metropolis–Hastings step (Carlin et al., 1992; Jacquier et al., 1994).

As noted by Carter and Kohn (1994), this sampler converges rather slowly when $\mathbf{Q}_t$ approaches singularity and breaks down to a reducible sampler; see also Pitt and Shephard (1999) for a theoretical investigation of this issue.

### Multi-Move Sampling of the Continuous State Process

A more efficient way to sample $\mathbf{x}_0, \dots, \mathbf{x}_T$ for the linear Gaussian state space model is joint or multi-move sampling of the states (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994; De Jong and Shephard, 1995; Koopman and Durbin, 2000). In contrast to single-move sampling, multi-move sampling draws the whole path $\mathbf{x} = (\mathbf{x}_0, \dots, \mathbf{x}_T)$ from the joint posterior of all states: $(\mathbf{x}_0, \dots, \mathbf{x}_T) \sim p(\mathbf{x}_0, \dots, \mathbf{x}_T|\mathbf{y})$. The multi-move sampler starts by representing the joint density $p(\mathbf{x}|\mathbf{y})$ as the product of $T + 1$ conditional densities:

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}_T|\mathbf{y}) \prod_{t=0}^{T-1} p(\mathbf{x}_t|\mathbf{x}_{t+1}, \dots, \mathbf{x}_T, \mathbf{y}). \tag{13.55}$$

The densities $p(\mathbf{x}_t|\mathbf{x}_{t+1}, \dots, \mathbf{x}_T, \mathbf{y})$ are the posterior densities of $\mathbf{x}_t$ knowing not only all observations $\mathbf{y}$, but also all future values $\mathbf{x}_{t+1}, \dots, \mathbf{x}_T$. This posterior is obtained by Bayes' theorem as

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y}) &\propto p(\mathbf{y}_{t+1},\dots,\mathbf{y}_T,\mathbf{x}_{t+1},\dots,\mathbf{x}_T|\mathbf{x}_t,\mathbf{y}^t)p(\mathbf{x}_t|\mathbf{y}^t)\\
&\propto \prod_{s=t+1}^{T} p(\mathbf{y}_s|\mathbf{x}_s)\prod_{s=t}^{T-1} p(\mathbf{x}_{s+1}|\mathbf{x}_s)p(\mathbf{x}_t|\mathbf{y}^t).
\end{aligned}
$$

Dropping terms that are independent of $\mathbf{x}_t$ we find that this density is obtained by combining the filter density $p(\mathbf{x}_t|\mathbf{y}^t)$ with the likelihood of $\mathbf{x}_{t+1}$ measured in terms of the transition density $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$:

$$
p(\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y}) \propto p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}^t). \tag{13.56}
$$

Equations (13.55) and (13.56) motivate was has been called forward-filtering-backward-sampling (Frühwirth-Schnatter, 1994).

*Algorithm 13.4: Forward-Filtering-Backward-Sampling (FFBS)*

(a) Determine and store the moments $\hat{\mathbf{x}}_{t|t}$ and $\mathbf{P}_{t|t}$ of the filtering density $p(\mathbf{x}_t|\mathbf{y}^t)$ by running a Kalman filter from $t = 1,\dots,T$ as described in *Algorithm 13.1.*
(b) Start sampling of the path $\mathbf{x}_0,\dots,\mathbf{x}_T$ by sampling the latest state vector $\mathbf{x}_T$ from the most recent filter density $p(\mathbf{x}_T|\mathbf{y}^T)$.
(c) Sample the remaining states $\mathbf{x}_t$ from $p(\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y})$ backward in time for $t = T-1,\dots,0$.

There exist various ways to implement step (c). Following Carter and Kohn (1994), $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ may be considered as the likelihood of a general linear model with known error covariance matrices as in Subsection 13.3.2, with observations $\mathbf{x}_{t+1}$ and regression parameter $\mathbf{x}_t$ following the conjugate normal prior $p(\mathbf{x}_t|\mathbf{y}^t)$. From Subsection 13.3.2, the density $p(\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y})$ is normal with

$$
\begin{aligned}
&\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y} \sim \mathcal{N}_d\left(\hat{\mathbf{x}}_{t|T}(\mathbf{x}_{t+1}),\mathbf{P}_{t|T}\right), \qquad (13.57)\\
&\hat{\mathbf{x}}_{t|T}(\mathbf{x}_{t+1}) = (\mathbf{I}-\mathbf{B}_{t+1}\mathbf{F}_{t+1})\hat{\mathbf{x}}_{t|t} + \mathbf{B}_{t+1}(\mathbf{x}_{t+1}-\mathbf{G}_{t+1}\mathbf{u}_{t+1}),\\
&\mathbf{P}_{t|T} = (\mathbf{I}-\mathbf{B}_{t+1}\mathbf{F}_{t+1})\mathbf{P}_{t|t},\\
&\mathbf{B}_{t+1} = \mathbf{P}_{t|t}\mathbf{F}_{t+1}^{'}\left(\mathbf{F}_{t+1}\mathbf{P}_{t|t}\mathbf{F}_{t+1}^{'} + \mathbf{Q}_{t+1}\right)^{-1}.
\end{aligned}
$$

If $\mathbf{Q}_{t+1}$ is positive definite, one could also use the information form of updating the posterior in a general linear model. If $\mathbf{Q}_{t+1}$ is singular, then the conditional density $p(\mathbf{x}_t|\mathbf{x}_{t+1},\dots,\mathbf{x}_T,\mathbf{y})$ is degenerate because part of $\mathbf{x}_t$ is deterministic given $\mathbf{x}_{t+1}$. Sampling from (13.57) based on a Cholesky decomposition of $\mathbf{P}_{t|T}$ will lead to numerical problems. Furthermore the recursions in (13.57) are inefficient, as they involve the inversion of a $(d \times d)$ matrix, with $d = \dim \mathbf{x}_t$, whereas $\mathbf{x}_t$ only has $s = \mathrm{rg}(\mathbf{Q}_t) < d$ random components. Frühwirth-Schnatter (1994) suggested transforming the state vector $\mathbf{x}_t$ to a new state variable with only $s$ random components. Another efficient sampler is to simulate the disturbances $\mathbf{w}_t$ rather than $\mathbf{x}_t$ using a disturbance smoother (De Jong and Shephard, 1995; Durbin and Koopman, 2002).

For more general state space models, such a multi-move sampler does not exist. Shephard and Pitt (1997) designed a blocked sampler, where an entire subblock $\mathbf{x}_t, \ldots, \mathbf{x}_{t+h}$ is sampled from the appropriate density using a Metropolis–Hastings step.

### 13.5.3 Sampling the Discrete States for a Switching State Space Model

The notation $\mathbf{S}^t = (S_0, \ldots, S_t)$ is used to denote a whole path of the hidden Markov chain $S_t$ up to $t$, with $S_0$ being dropped for finite mixtures of state space models.

#### Full Conditional Sampling of a Hidden Markov Chain

Full conditional sampling of the states $\mathbf{S}$ of a hidden Markov chain is not restricted to linear Gaussian state space models, but may be applied also to more general models with nonnormal or nonlinear densities $p(\mathbf{y}_t|S_t = j, \mathbf{x}_t, \boldsymbol{\vartheta})$ and $p(\mathbf{x}_t|S_t = j, \mathbf{x}_{t-1}, \boldsymbol{\vartheta})$.

Single-move sampling of $p(S_t|\mathbf{S}_{-t}, \mathbf{x}, \mathbf{y}, \boldsymbol{\vartheta})$ could be used as in Subsection 11.5.6, however, it is much more efficient to use a multi-move sampler (Carter and Kohn, 1994; Shephard, 1994) that samples the whole path $\mathbf{S} = (S_0, \ldots, S_T)$ jointly from $p(\mathbf{S}|\mathbf{x}, \mathbf{y}, \boldsymbol{\vartheta})$. This multi-move sampler is closely related to the sampler discussed in *Algorithm 11.5* for finite Markov mixture models.

*Algorithm 13.5: Multi-Move Sampling of the Discrete States of a Switching State Space Model*

(a) Run a filter conditional on $\boldsymbol{\vartheta}$ and $\mathbf{x}$ to obtain the filtered probability distribution $\Pr(S_t = j|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta})$ for $t = 1, \ldots, T$. The filter is started at $t = 1$ with the initial distribution $\Pr(S_0 = k|\boldsymbol{\xi})$. For each $t \geq 1$, perform one-step ahead prediction,

$$\Pr(S_t = j|\mathbf{y}^{t-1}, \mathbf{x}^{t-1}, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} \xi_{kj} \Pr(S_{t-1} = k|\mathbf{y}^{t-1}, \mathbf{x}^{t-1}, \boldsymbol{\vartheta}),$$

and filtering for each possible value $j = 1, \ldots, K$ of $S_t$:

$$\begin{aligned} &\Pr(S_t = j|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta}) \\ &\propto p(\mathbf{y}_t|S_t = j, \mathbf{x}_t, \boldsymbol{\vartheta}) p(\mathbf{x}_t|S_t = j, \mathbf{x}_{t-1}, \boldsymbol{\vartheta}) \Pr(S_t = j|\mathbf{y}^{t-1}, \mathbf{x}^{t-1}, \boldsymbol{\vartheta}). \end{aligned} \qquad (13.58)$$

The probabilities in (13.58) need to be normalized to obtain a proper filter distribution.

(b) Sample $S_T$ from the discrete probability distribution $\Pr(S_T = j|\mathbf{y}^T, \mathbf{x}^T, \boldsymbol{\vartheta})$.

(c) For $t = T-1, T-2, \ldots, 0$ sample $S_t$ from the conditional distribution $\Pr(S_t = j|S_{t+1}, \mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta})$ given by

$$\Pr(S_t = j|S_{t+1}, \mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta}) = \frac{\xi_{j,S_{t+1}} \Pr(S_t = j|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta})}{\sum_{k=1}^{K} \xi_{k,S_{t+1}} \Pr(S_t = k|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta})}.$$

Here $S_{t+1}$ is the most recent value sampled for the hidden Markov chain at $t+1$.

### Marginal Sampling of the Indicators

Both Carter and Kohn (1996) and Gerlach and Kohn (2000) generate $S_t$ from the discrete density $p(S_t|\mathbf{S}_{-t}, \mathbf{y}, \boldsymbol{\vartheta})$ without conditioning on the continuous states $\mathbf{x}$. Marginalization over $\mathbf{x}$, however, leads to dependence among all the values of $S_t$, even if the indicators are i.i.d., and generating $S_t$ in an efficient way is far from straightforward.

Suppose that $\mathbf{S}^{t-1}$ has already been updated and that the first two moments of the normal density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1}, \mathbf{S}^{t-1}, \boldsymbol{\vartheta})$ are known. Bayes' theorem is used to obtain the density $p(S_t|\mathbf{S}_{-t}, \mathbf{y}, \boldsymbol{\vartheta})$:

$$p(S_t|\mathbf{S}_{-t}, \mathbf{y}, \boldsymbol{\vartheta}) \propto p(S_t|\mathbf{S}_{-t}, \boldsymbol{\vartheta}) p(\mathbf{y}_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta}) p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\mathbf{y}^t, \mathbf{S}_{-t}, S_t, \boldsymbol{\vartheta}).$$

For each of the $K$ values of $S_t$, the predictive density $p(\mathbf{y}_t|\mathbf{y}^{t-1}, \mathbf{S}^t, \boldsymbol{\vartheta})$ as well as the filtering density $p(\mathbf{x}_t|\mathbf{y}^t, \mathbf{S}^t, \boldsymbol{\vartheta})$, is obtained from a single step of the Kalman filter. A direct but inefficient method to evaluate the predictive density $p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\mathbf{y}^t, \mathbf{S}_{-t}, S_t)$ for the $K$ different values of $S_t$ is to use $T-t+1$ forecasting steps of the Kalman filter, which requires $\mathcal{O}(T)$ steps to generate $S_t$, and hence $\mathcal{O}(T^2)$ steps to generate the whole path $\mathbf{S}$. Gerlach and Kohn (2000) show how to obtain the term $p(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T|\mathbf{y}^t, \mathbf{S}_{-t}, S_t)$ in one step after an initial set of backward recursions, requiring $\mathcal{O}(T)$ steps to generate the whole path $\mathbf{S}$. We refer to Gerlach and Kohn (2000) for more details.

Finally, Gerlach and Kohn (2000) discuss an efficient way of sampling a binary indicator $S_t$ which takes one of two values most of the time, for instance, an indicator corresponding to an outlier or to an intervention variable.

## 13.6 Further Issues

### 13.6.1 Model Specification Uncertainty in Switching State Space Modeling

The application of the state space approach to socioeconomic or biological sciences is complicated by the need of model identification, because often little a priori information about the dynamics of the system is available. One approach toward this model specification uncertainty is to fit several state space models to a given time series and to apply some method of model selection.

AIC was used in the context of model selection for state space models by, among others, Kitagawa (1981) and Harvey (1989). AIC and BIC are defined for state space models in the usual way as

$$\mathrm{AIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}) + 2 \dim(\boldsymbol{\vartheta}), \tag{13.59}$$

$$\mathrm{BIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}) + \log(T) \dim(\boldsymbol{\vartheta}), \tag{13.60}$$

where $p(\mathbf{y}|\hat{\boldsymbol{\vartheta}})$ is the (approximate) likelihood of a (switching) state space model evaluated at the ML estimator $\hat{\boldsymbol{\vartheta}}$. Durbin and Koopman (2001, p.152) provide a corrected AIC and BIC for state space models with diffuse initial conditions. Harvey (1989) and Durbin and Koopman (2001) prefer a definition where the right-hand side of (13.59) and (13.60) is divided by $T$.

The marginal likelihood has been applied to model selection problems involving state space models by, among many others, Frühwirth-Schnatter (1995), Shively and Kohn (1997), and Koop and van Dijk (2000). Frühwirth-Schnatter (2001a) discusses model comparison based on marginal likelihoods for switching linear Gaussian state space models and uses the bridge sampling techniques discussed in Subsection 5.4.6 to obtain a numerical approximation of the marginal likelihood.

A Bayesian variable selection approach (Carlin and Chib, 1995) has been applied to switching dynamic factor models by Kim and Nelson (2001).

### 13.6.2 Auxiliary Mixture Sampling for Nonlinear and Nonnormal State Space Models

To deal with non-Gaussian or nonlinear state space models it is useful to approximate nonnormal densities by a finite mixture of common distributions. Sorenson and Alspach (1971) and Alspach and Sorenson (1972) are pioneering works using a Gaussian sum approximation to derive an approximate filter for nonlinear and non-Gaussian state space models. Meinhold and Singpurwalla (1989) represent the posterior density $p(\mathbf{x}_{t-1}|\mathbf{y}^{t-1})$ by a mixture of $t$-distributions and suggest some approximate recursive scheme to obtain a similar mixture approximation to $p(\mathbf{x}_t|\mathbf{y}^t)$.

To facilitate statistical inference, Shephard (1994) introduced the concept of partially Gaussian state space models and suggested approximating nonnormal densities appearing in the definition of the state space model by mixtures of normal distributions. This allows MCMC estimation through efficient multi-move sampling of the state process as in *Algorithm 13.4* also for non-Gaussian state space models, where usually single-move sampling has to be applied.

MCMC methods based on a finite mixture approximation have been developed in particular for stochastic volatility models (Shephard, 1994; Kim et al., 1998; Chib et al., 2002; Omori et al., 2004). A stochastic volatility model is a state space model with state vector $h_t$, usually assumed to follow an AR(1)-process, where the observation equation is nonlinear, because the variance of the observation error is a nonlinear function of $h_t$:

$$\begin{aligned} h_t &= \delta h_{t-1} + \zeta + w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \\ Y_t &= e^{h_t/2} z_t, \qquad z_t \sim \mathcal{N}\left(0, 1\right). \end{aligned}$$

This model may be transformed into a linear state space model with nonnormal errors in the following way,

$$\log Y_t^2 = h_t + \varepsilon_t,$$

where $\varepsilon_t$ is equal to the log of a $\chi_1^2$ random variable. The density of the $\log \chi_1^2$ is approximated in Shephard (1994) by a mixture of univariate normal distributions,

$$p(\varepsilon_t) = \sum_{k=1}^{K} w_k f_N(\varepsilon_t; m_k, s_k^2).$$

Shephard (1994) derived appropriate parameters $(w_k, m_k, s_k^2), k = 1, \ldots, K$, for mixtures up to $K = 7$ components, whereas a more accurate approximation with $K = 10$ components appears in Omori et al. (2004). By introducing i.i.d. hidden indicators $S_t$ for each $t$, the following finite mixture of linear Gaussian state space models results,

$$\begin{aligned} h_t &= \delta h_{t-1} + \zeta + w_t, \qquad w_t \sim \mathcal{N}\left(0, \sigma_\mu^2\right), \\ \log Y_t^2 &= h_t + m_{S_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, s_{S_t}^2\right), \end{aligned}$$

with $\Pr(S_t = k) = w_k$. Filtering and parameter estimation as discussed in Sections 13.3 to 13.5 may be applied.

Recently, Frühwirth-Schnatter and Wagner (2006) developed a similar auxiliary mixture sampler for state space modeling of count data, based on a finite mixture approximation to the type I extreme value distribution. Frühwirth-Schnatter and Frühwirth (2006) show that this sampler may be extended to deal with state space modeling of binary and multinomial data.

## 13.7 Illustrative Application to Modeling Exchange Rate Data

For illustration we reanalyze the U.S./U.K. real exchange rate from January 1885 to November 1995, originally published in Grilli and Kaminsky (1991)

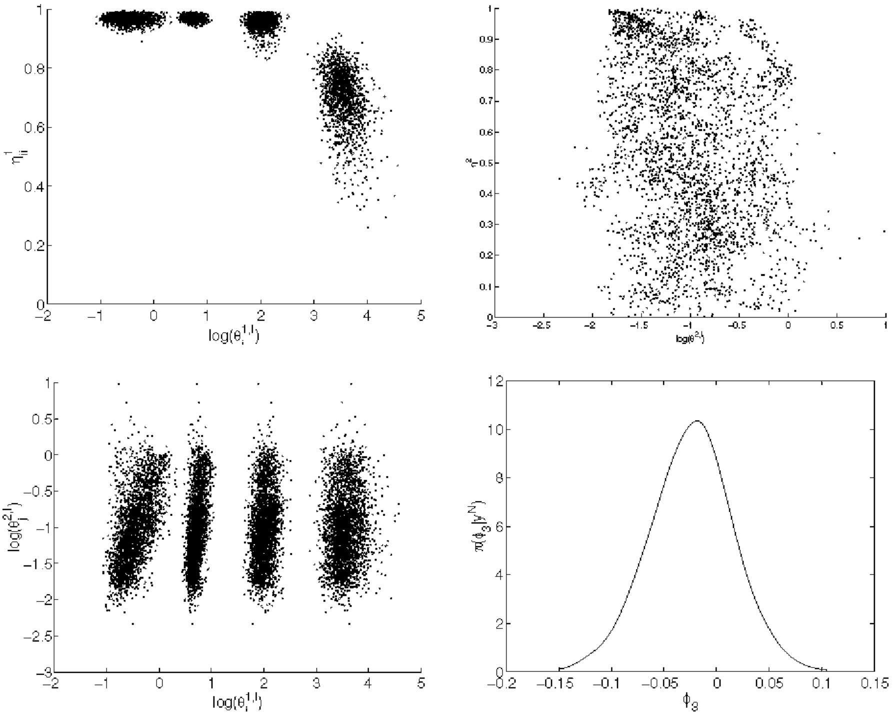

**Fig. 13.1.** U.S./U.K. Real Exchange Rate Data, exploratory Bayesian analysis for a switching model with $K_1 = 4, K_2 = 2, p = 3$; top left-hand side: $\log(\sigma^2_{1,k})$ versus $\boldsymbol{\xi}^1_{kk}$ for all possible $k$; top right-hand side: $\log(\sigma^2_{2,k})$ versus $\boldsymbol{\xi}^2_{kk}$ for all possible $k$; bottom left-hand side: $\log(\sigma^2_{1,k})$ versus $\log(\sigma^2_{2,k})$ for all possible $k$; bottom right-hand side: posterior of $\delta_3$ (from Frühwirth-Schnatter (2001a) with permission granted by The Institute of Statistical Mathematics)

and reanalyzed by Engel and Kim (1999) and Frühwirth-Schnatter (2001a). The real exchange rate is defined as the relative price of U.K. to U.S. producer goods; that is, U.S./U.K. nominal exchange rate times the U.K. producer price index divided by the U.S. producer price index. Engel and Kim (1999) suggested decomposing the log of the real exchange rate $Y_t$ into a permanent component $\mu_t$ and a transitory component $c_t$:

$$\log Y_t = \mu_t + c_t,$$

where $c_t$ is assumed to follow an AR($p$) process:

$$c_t = \delta_1 c_{t-1} + \cdots + \delta_p c_{t-p} + w_{t,1},$$

and $\mu_t$ follows a random walk process:

$$\mu_t = \mu_{t-1} + w_{t,2}.$$

The conditional variance of the transitory component $c_t$ is assumed to switch between $K_1$ values according to a Markov chain $S_t^1$ with transition matrix $\boldsymbol{\xi}^1$, whereas the conditional variance of the permanent component $\mu_t$ is assumed to switch between $K_2$ values according to a Markov chain $S_t^2$ with transition matrix $\boldsymbol{\xi}^2$:

$$w_{t,1} \sim \mathcal{N}\left(0, \sigma^2_{1,S_t^1}\right), \qquad w_{t,2} \sim \mathcal{N}\left(0, \sigma^2_{2,S_t^2}\right).$$

The model can be put into state space form with the following state vector $\mathbf{x}_t$ and matrix $\mathbf{F}$,

$$\mathbf{x}_t = \begin{pmatrix} \mu_t \\ c_t \\ \vdots \\ c_{t-p+1} \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathbf{F}(\boldsymbol{\delta}) \end{pmatrix},$$

and $\mathbf{F}(\boldsymbol{\delta})$ being the same as in (13.15).

This model is a switching linear Gaussian state space model with two hidden indicators. The estimation method used by Frühwirth-Schnatter (2001a) is an extension of *Algorithm 13.2* to the case of two hidden switching variables. Frühwirth-Schnatter (2001a) did not condition on the first values of the state process as in Engel and Kim (1999), but sample in step (a) the whole processes $c_{1-p}, \dots, c_0, \dots, c_T$ and $\mu_0, \dots, \mu_T$ including the starting values by applying the multi-move sampler of Frühwirth-Schnatter (1994). The filter is initialized with the prior $\mathbf{x}_0 \sim \mathcal{N}\left(\hat{\mathbf{x}}_{0|0}, \mathbf{P}_{0|0}\right)$, where

$$\hat{\mathbf{x}}_{0|0} = \begin{pmatrix} \log y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \mathbf{P}_{0|0} = \begin{pmatrix} 1000 & \mathbf{0}_{1\times(p-1)} \\ \mathbf{0}_{(p-1)\times 1} & \mathbf{M} \end{pmatrix}$$

with

$$\mathrm{vec}(\mathbf{M}) = (\mathbf{I}_{p^2} - \mathbf{F}(\boldsymbol{\delta}) \otimes \mathbf{F}(\boldsymbol{\delta}))^{-1} \begin{pmatrix} \sigma^2_{1,S_0^1} \\ \mathbf{0}_{(p-1)\times 1} \end{pmatrix},$$

and $\otimes$ is the Kronecker product of two matrices. This choice is based on the suggestion of De Jong and Chu-Chun-Lin (1994) for combining a vague prior with a stationary prior for state vectors containing both nonstationary and stationary components.

As the Markov processes $S_t^1$ and $S_t^2$ are independent a posteriori, sampling in step (b) is carried out independently for both indicators using *Algorithm 13.5*. For $S_t^1$ the filter step (13.58) is based on

$$\begin{aligned} &\Pr(S_t^1 = j|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta}) \\ &\propto f_N(c_t; \delta_1 c_{t-1} + \cdots + \delta_p c_{t-p}, \sigma^2_{1,j})\Pr(S_t^1 = j|\mathbf{y}^{t-1}, \mathbf{x}^{t-1}, \boldsymbol{\vartheta}), \end{aligned}$$

whereas for $S_t^2$ this step reads:

$$\Pr(S_t^2 = j|\mathbf{y}^t, \mathbf{x}^t, \boldsymbol{\vartheta}) \propto f_N(\mu_t; \mu_{t-1}, \sigma_{2,j}^2)\Pr(S_t^2 = j|\mathbf{y}^{t-1}, \mathbf{x}^{t-1}, \boldsymbol{\vartheta}).$$

Parameter estimation is based on the priors $\sigma_{1,k}^2 \sim \mathcal{G}^{-1}(3, 8)$, $k = 1, \ldots, K_1$, and $\sigma_{2,k}^2 \sim \mathcal{G}^{-1}(3, 2)$, $k = 1, \ldots, K_2$. The prior for all rows of the transition matrices $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ is chosen to be $\mathcal{D}(1, \ldots, 1)$.

All variances $\sigma_{1,k}^2, k = 1, \ldots, K_1$ and $\sigma_{2,k}^2, k = 1, \ldots, K_2$ are sampled at the same time, as they are conditionally independent, inverted Gamma distributed. This is different from Engel and Kim (1999) who impose a priori an identifiability constraint on the variances and sample the variances in a single-move manner from the constrained posterior.

Sampling of the AR($p$) parameters $\delta_1, \ldots, \delta_p$ is carried out from the regression model $c_t = \delta_1 c_{t-1} + \cdots + \delta_p c_{t-p} + \sigma_{1,S_t^1}\varepsilon_t$, where $\varepsilon_t$ is i.i.d. standard normal. As samples of $c_0, \ldots, c_{1-r}$ are available from step (a), $t$ is running from 1 to $T$. Within one iteration, sampling of the AR($p$) parameters $\delta_1, \ldots, \delta_p$ is repeated until the stationarity condition on the AR($p$) process is fulfilled.

**Table 13.1.** U.S./U.K. Real Exchange Rate Data, model selection using marginal likelihoods (from Frühwirth-Schnatter (2001a) with permission granted by The Institute of Statistical Mathematics)

| Model | $\log p(\mathbf{y}\|\text{Model})$ |
|---|---|
| $K_1 = 4$, $K_2 = 2$, $p = 3$ | –2562.4 |
| $K_1 = 4$, $K_2 = 1$, $p = 2$ | **–2515.5** |
| $K_1 = 4$, $K_2 = 1$, $p = 1$ | –2612.5 |
| $K_1 = 3$, $K_2 = 1$, $p = 2$ | –2605.9 |
| $K_1 = 5$, $K_2 = 1$, $p = 2$ | –2880.2 |
| No switching, $p = 2$ | –2914.4 |

Engel and Kim (1999) selected a model where the variance of the transitory component is driven by a three-state Markov switching process, the variance of the permanent component is constant, and the order of the AR process is equal to two, that is, $K_1 = 3, K_2 = 1$, $p = 2$. They adopt this specification by exploring the posterior distributions without formal Bayesian model selection.

We proceed with an exploratory Bayesian analysis of a model with $K_1 = 4$, $K_2 = 2$, and $p = 3$, using the MCMC output of a random permutation sampler. Parts (a) and (b) of Figure 13.1 show a point process representation of $(\sigma_{1,k}^2)^{(m)}$ versus $(\boldsymbol{\xi}_{kk}^1)^{(m)}$ and $(\sigma_{2,j}^2)^{(m)}$ versus $(\boldsymbol{\xi}_{jj}^2)^{(m)}$ for all possible states $k \in \{1, \ldots, K_1\}$ and $j \in \{1, \ldots, K_2\}$, respectively. For $S_t^1$ we have allowed for four states and there are actually four simulation clusters; for $S_t^2$, however, we have allowed for two states, but there is just one simulation cluster. This provides empirical evidence in favor of a homogeneous rather than a switching variance of the permanent component. This hypothesis is further supported by part (c) of the figure where the point process representation of $(\sigma^2_{1,k})^{(m)}$ versus $(\sigma^2_{2,k})^{(m)}$ is plotted. Finally, part (d) of the same figure plots the posterior of the AR parameter $\delta_3$ which may be estimated directly from the output of the random permutation sampler as $\delta_3$ is state independent. The mode of the posterior is close to 0 providing evidence for the hypothesis that $\delta_3$ is equal to zero. To sum up, the exploratory analysis provides evidence in favor of a model with $K_1 = 4$, $K_2 = 1$, and $p = 2$ rather than $K_1 = 3$, $K_2 = 2$, and $p = 2$.

In Frühwirth-Schnatter (2001a) the marginal likelihood, based on a bridge sampling estimator, was used for model selection; see Table 13.1. For the best model the variance of the transitory component is driven by a four-state Markov switching process, the variance of permanent component is constant, and the order of the AR process is equal to two; that is, $K_1 = 4$, $K_2 = 1$, $p = 2$.

The marginal likelihoods reported in Table 13.1, however, clearly favor the model with $K_1 = 4$, $K_2 = 1$, and $p = 2$, which differs from the one selected in Engel and Kim (1999) by the number of states of the variance of the transitory component. Increasing the number of states from four to five, however, reduces the marginal likelihood drastically. For completeness, the marginal likelihood for a model without switching is reported, showing that this model is the most unlikely of all.

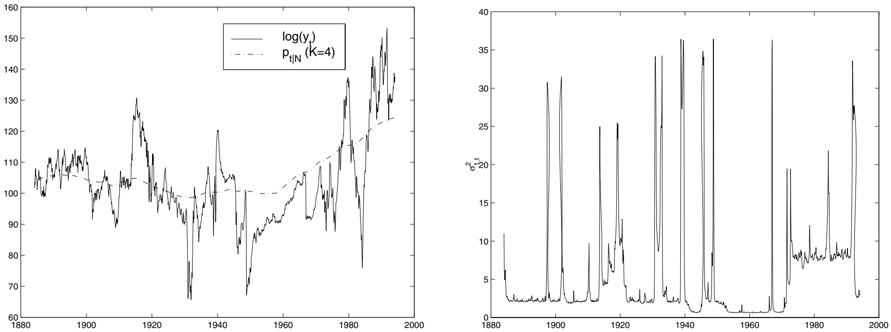

**Fig. 13.2.** U.S./U.K. Real Exchange Rate Data, four-state model ($K_1 = 4$, $K_2 = 1$, $p = 2$); left-hand side: smoothed real exchange rate $\hat{p}_{t|T}$; right-hand side: estimated time-varying variance $\hat{\sigma}^2_{1,t}$ ($K_1 = 3$, $K_2 = 1$, $p = 2$) (from Frühwirth-Schnatter (2001a) with permission granted by The Institute of Statistical Mathematics)

We can draw further interesting inferences from the output of the random permutation sampler without the need to identify the model. This is especially true for the smoothed permanent component $\hat{p}_{t|T}$ which is compared in Figure 13.2 with the observed time series. The resulting estimator of the permanent component is much smoother than the rather noisy estimate published in Engel and Kim (1999), being nearly constant until the end of the fifties and increasing afterwards. Another interesting picture is obtained if we plot the time-varying variance $\sigma^2_{1,t}$ estimated from:

$$\hat{\sigma}^2_{1,t} = \frac{1}{M} \sum_{m=1}^{M} \left(\sigma^2_{1,s}\right)^{(m)},$$

where $s = (S^1_t)^{(m)}$ over time $t$ as in Figure 13.2.

**Table 13.2.** U.S./U.K. Real Exchange Rate Data, estimation results for $K_1 = 4, K_2 = 1, p = 2$ (from Frühwirth-Schnatter (2001a) with permission granted by The Institute of Statistical Mathematics)

| Parameter | Mean | Std.Dev. | 95%-H.P.D. Regions | |
|---|---|---|---|---|
| $\sigma^2_{1,1}$ | 0.634 | 0.151 | 0.371 | 0.93 |
| $\sigma^2_{1,2}$ | 2.05 | 0.196 | 1.67 | 2.42 |
| $\sigma^2_{1,3}$ | 7.63 | 1.07 | 5.9 | 9.88 |
| $\sigma^2_{1,4}$ | 36.4 | 9.13 | 20.7 | 53.9 |
| $\sigma^2_2$ | 0.366 | 0.132 | 0.121 | 0.608 |
| $\delta_1$ | 1.06 | 0.0474 | 0.967 | 1.14 |
| $\delta_2$ | –0.0729 | 0.046 | –0.158 | 0.0139 |
| $\xi_{11}$ | 0.968 | 0.0132 | 0.943 | 0.991 |
| $\xi_{12}$ | 0.0091 | 0.00861 | 2.84e–006 | 0.0256 |
| $\xi_{13}$ | 0.00639 | 0.00586 | 2.87e–006 | 0.0189 |
| $\xi_{14}$ | 0.0162 | 0.00987 | 0.000231 | 0.0341 |
| $\xi_{21}$ | 0.00855 | 0.00576 | 0.000165 | 0.0205 |
| $\xi_{22}$ | 0.973 | 0.00853 | 0.957 | 0.988 |
| $\xi_{23}$ | 0.00587 | 0.0057 | 6.19e–006 | 0.0155 |
| $\xi_{24}$ | 0.0123 | 0.00697 | 0.000484 | 0.0246 |
| $\xi_{31}$ | 0.00498 | 0.00489 | 1.24e–005 | 0.0144 |
| $\xi_{32}$ | 0.0139 | 0.0123 | 9.59e–006 | 0.0373 |
| $\xi_{33}$ | 0.956 | 0.0222 | 0.916 | 0.992 |
| $\xi_{34}$ | 0.0248 | 0.0161 | 0.00129 | 0.0562 |
| $\xi_{41}$ | 0.039 | 0.0338 | 0.000159 | 0.103 |
| $\xi_{42}$ | 0.147 | 0.0691 | 0.024 | 0.288 |
| $\xi_{43}$ | 0.123 | 0.0934 | 0.00108 | 0.309 |
| $\xi_{44}$ | 0.691 | 0.116 | 0.438 | 0.865 |

The selected model has to be identified to draw inference on the variances of the different states as well as to obtain state estimates over the whole observation period. The identifiability constraint $\sigma^2_{1,1} < \sigma^2_{1,2} < \sigma^2_{1,3} < \sigma^2_{1,4}$ is suggested by the point process representation in Figure 13.1, showing that the states of $S^1_t$ differ in the variance of the transitory component. If this constraint is included in the permutation sampler, no label switching occurs.

Table 13.2 reports point estimates as well as 95%-H.P.D.-regions for all model parameters, including estimates of the state-specific variances as well as estimates of the transition probabilities.

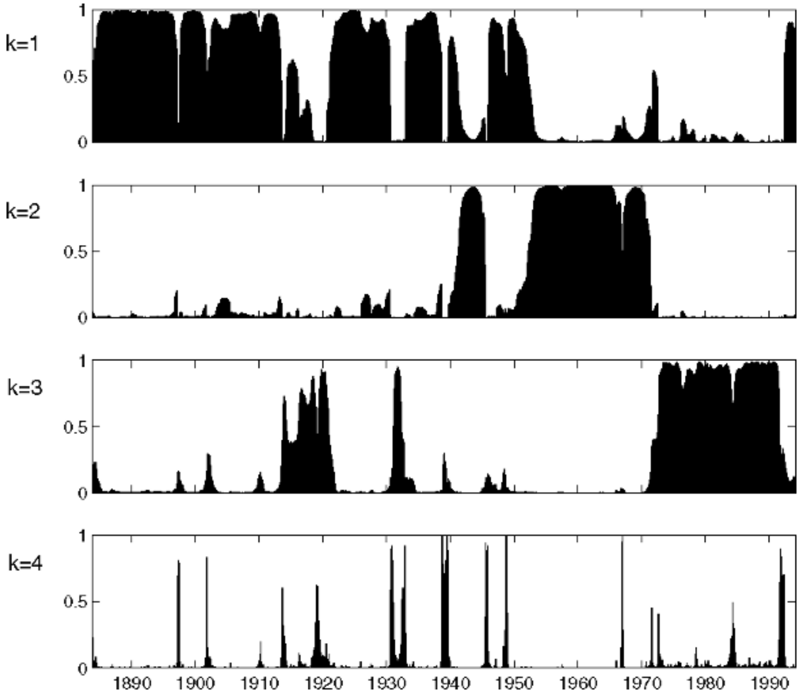

**Fig. 13.3.** U.S./U.K. Real Exchange Rate Data, smoothed state probabilities for $S_t^1$ for a switching state space model with $K_1 = 4$, $K_2 = 1$, and $p = 2$ (from Frühwirth-Schnatter (2001a) with permission granted by The Institute of Statistical Mathematics)

Figure 13.3 plots the smoothed posterior state probabilities $\Pr(S_t^{1,\mathcal{L}} = k|\mathbf{y})$ of being in a certain state $k \in \{1,2,3,4\}$ over time $t$, for a four-state switching model, and compares them with the probabilities obtained from the three-state model. The probabilities $\Pr(S_t^{1,\mathcal{L}} = k|\mathbf{y})$ are estimated from the constrained MCMC output by

$$\Pr(S_t^{1,\mathcal{L}} = k|\mathbf{y}) = \frac{1}{M}\#\{(S_t^{1,\mathcal{L}})^{(m)} = k\}.$$

Engel and Kim (1999) found the following interpretation of these probabilities. The quietest state occurred during the first half of the forties and then from about 1952 to the end of the seventies, which are periods in which the nominal exchange rate was fixed. The two medium-state variances correspond to periods of floating nominal exchange rates. Periods of high-state variance are rather singular events and can be identified with specific historical events.