# 7

# Bayesian Estimation of the Markov-Switching GJR(1, 1) Model with Student-$t$ Innovations

> *(...) "the application of GARCH to long time series of stock-return data will yield a high measure of persistence because of the presence of deterministic shifts in the unconditional variance and the subsequent failure of the econometrician to model these shifts."*
>
> — *Christopher G. Lamoureux and William D. Lastrapes*

In this chapter, we address the problem of estimating GARCH models subject to structural changes in the parameters; namely, the Markov-switching GARCH models (henceforth MS-GARCH). In this framework, a hidden Markov sequence $\{s_t\}$ with state space $\{1, \ldots, K\}$ allows discrete changes in the model parameters. Such processes have received a lot of attention in recent years as they provide an explanation of the high persistence in volatility (*i.e.*, nearly unit root process for the conditional variance) observed with single-regime GARCH models [see, *e.g.*, Lamoureux and Lastrapes 1990]. Furthermore, the MS-GARCH models allow for a quick change in the volatility level which leads to significant improvements in volatility forecasts, as shown by Dueker [1997], Klaassen [2002], Marcucci [2005].

Following the seminal work of Hamilton and Susmel [1994], different parametrizations have been proposed to account for discrete changes in the GARCH parameters [see, *e.g.*, Dueker 1997, Gray 1996, Klaassen 2002]. However, these parametrizations for the conditional variance process lead to computational difficulties. Indeed, the evaluation of the likelihood function for a sample of length $T$ requires the integration over all $K^T$ possible paths, rendering the estimation infeasible. As a remedy, approximation schemes have been proposed to shorten the dependence on the state variable's history. While this difficulty is not present in ARCH type models, lower order GARCH specification of the conditional variance offers a more parsimonious representation than higher order ARCH models.

In order to avoid any difficulties related to the past infinite history of the state variable, we adopt a recent parametrization due to Haas *et al.* [2004]. In their model, the authors hypothesize $K$ separate GARCH$(1,1)$ processes for the conditional variance of the MS-GARCH process $\{y_t\}$. The conditional variances at time $t$ can be written in vector form as follows:

$$\begin{pmatrix} h_t^1 \\ h_t^2 \\ \vdots \\ h_t^K \end{pmatrix} \doteq \begin{pmatrix} \alpha_0^1 \\ \alpha_0^2 \\ \vdots \\ \alpha_0^K \end{pmatrix} + \begin{pmatrix} \alpha_1^1 \\ \alpha_1^2 \\ \vdots \\ \alpha_1^K \end{pmatrix} y_{t-1}^2 + \begin{pmatrix} \beta^1 \\ \beta^2 \\ \vdots \\ \beta^K \end{pmatrix} \odot \begin{pmatrix} h_{t-1}^1 \\ h_{t-1}^2 \\ \vdots \\ h_{t-1}^K \end{pmatrix} \tag{7.1}$$

where $\odot$ denotes the Hadamard product, *i.e.*, element-by-element multiplication. The MS-GARCH process $\{y_t\}$ is then simply obtained by setting:

$$y_t = \varepsilon_t (h_t^{s_t})^{1/2}$$

where $\varepsilon_t$ is an error term with zero mean and unit variance. The parameters $\alpha_0^k$, $\alpha_1^k$ and $\beta^k$ are therefore the GARCH$(1,1)$ parameters related to the $k$th state of the nature. Under this specification, the conditional variance is solely a function of the past data and current state $s_t$, which avoids the problem of infinite history. In the context of the Bayesian estimation, this allows to simulate the state process in a multi-move manner which enhances the sampler's efficiency.

In addition to its appealing computational aspects, the MS-GARCH model of Haas *et al.* [2004] has conceptual advantages. In effect, one reason for specifying Markov-switching models that allow for different GARCH behavior in each regime is to capture the difference in the variance dynamics in low- and high-volatility periods. As pointed out by Haas *et al.* [2004, p.498]:

> (...) "a relatively large value of $\alpha_1^k$ and relatively low values of $\beta^k$ in high-volatility regimes may indicate a tendency to over-react to news, compared to *regular* periods, while there is less memory in these subprocesses. Such an interpretation requires a parametrization of Markov-switching GARCH models that implies a clear association between the GARCH parameters within regime $k$, that is $\alpha_0^k$, $\alpha_1^k$ and $\beta^k$ and the corresponding $\{h_t^k\}$ process."

The specification of the conditional variance in equation (7.1) allows for a clear-cut interpretation of the variance dynamics in each regime. Moreover, Haas *et al.* [2004] show that results on the single-regime GARCH$(1,1)$ model can be extended to their specification; in particular, they derive explicit formulae for the covariance stationarity condition, the unconditional variance as well as the dependence structure of the squared process $\{y_t^2\}$.

To account for additional stylized facts observed in financial time series, especially for stock indices (see **Chap. 4**), we will consider an asymmetric extension of (7.1) in which the GARCH$(1,1)$ processes are replaced by GJR$(1,1)$ processes. More precisely, in this Markov-switching GJR model (henceforth MS-GJR), the conditional variances at time $t$ can be written in vector form as follows:

$$\begin{pmatrix} h_t^1 \\ h_t^2 \\ \vdots \\ h_t^K \end{pmatrix} \doteq \begin{pmatrix} \alpha_0^1 \\ \alpha_0^2 \\ \vdots \\ \alpha_0^K \end{pmatrix} + \left[ \begin{pmatrix} \alpha_1^1 \\ \alpha_1^2 \\ \vdots \\ \alpha_1^K \end{pmatrix} \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \begin{pmatrix} \alpha_2^1 \\ \alpha_2^2 \\ \vdots \\ \alpha_2^K \end{pmatrix} \mathbb{I}_{\{y_{t-1} < 0\}} \right] y_{t-1}^2 + \begin{pmatrix} \beta^1 \\ \beta^2 \\ \vdots \\ \beta^K \end{pmatrix} \odot \begin{pmatrix} h_{t-1}^1 \\ h_{t-1}^2 \\ \vdots \\ h_{t-1}^K \end{pmatrix} \tag{7.2}$$

where $\mathbb{I}_{\{\bullet\}}$ denotes the indicator function. In this setting, the conditional variance in every regime can react asymmetrically depending on the sign of the past shocks due to the introduction of dummy variables. The leverage effect is present for a given state $k$ as soon as $\alpha_2^k > \alpha_1^k$. An interesting feature of the parametrization (7.2) lies in the fact that we can estimate whether the response to past negative shock on the conditional variance is different across regimes.

The plan of this chapter is as follows. We set up the model in **Sect. 7.1**. The MCMC scheme is detailed in **Sect. 7.2**. The MS-GJR model as well as a single-regime GJR model are applied to the Swiss Market Index log-returns in **Sect. 7.3**. In **Sect. 7.4**, we test the models for misspecification by using the generalized residuals and assess the goodness-of-fit through the calculation of the Deviance information criterion and the model likelihoods. In **Sect. 7.5**, we test the predictive performance of the models by running a forecasting analysis based on the VaR. In **Sect. 7.6**, we propose a methodology to depict the one-day ahead VaR density and document how specific forecasters' risk perspectives can lead to different conclusions in terms of the forecasting performance of the model. We conclude with some comments regarding the ML estimation of the MS-GJR model in **Sect. 7.7**.

## 7.1 The model and the priors

A Markov-switching GJR$(1,1)$ model with Student-$t$ innovations may be written as follows:

$$\begin{aligned} y_t &= \varepsilon_t(\varrho h_t)^{1/2} \quad \text{for } t = 1, \dots, T \\ \varepsilon_t &\overset{iid}{\sim} \mathcal{S}(0, 1, \nu) \\ \varrho &\doteq \frac{\nu - 2}{\nu} \\ h_t &\doteq \mathbf{e}_t' \mathbf{h}_t \end{aligned} \tag{7.3}$$

where $\mathbf{e}_t$ is a $K \times 1$ vector defined by $\mathbf{e}_t \doteq \left(\mathbb{I}_{\{s_t=1\}} \cdots \mathbb{I}_{\{s_t=K\}}\right)'$, $\mathbb{I}_{\{\bullet\}}$ is the indicator function; the sequence $\{s_t\}$ is assumed to be a stationary, irreducible Markov process with discrete state space $\{1, \dots, K\}$ and transition matrix $P \doteq [P_{ij}]$ where $P_{ij} \doteq \mathbb{P}(s_{t+1} = j \mid s_t = i)$; $\mathcal{S}(0, 1, \nu)$ denotes the standard Student-$t$ density with $\nu$ degrees of freedom and $\varrho$ is a scaling factor which ensures that the conditional variance is given by $h_t$. Moreover, we define the $K \times 1$ vector of GJR$(1, 1)$ conditional variances in a compact form as follows:

$$\mathbf{h}_t \doteq \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1} < 0\}}) y_{t-1}^2 + \boldsymbol{\beta} \odot \mathbf{h}_{t-1}$$

where $\mathbf{h}_t \doteq (h_t^1 \cdots h_t^K)'$, $\boldsymbol{\alpha}_j \doteq (\alpha_j^1 \cdots \alpha_j^K)'$ for $j = 0, 1, 2$ and $\boldsymbol{\beta} \doteq (\beta^1 \cdots \beta^K)'$. In addition, we require that $\boldsymbol{\alpha}_0 > \mathbf{0}$, $\boldsymbol{\alpha}_1 \geqslant \mathbf{0}$, $\boldsymbol{\alpha}_2 \geqslant \mathbf{0}$ and $\boldsymbol{\beta} \geqslant \mathbf{0}$, where $\mathbf{0}$ is a $K \times 1$ vector of zeros, in order to ensure the positivity of the conditional variance in every regime and set $\mathbf{h}_0 \doteq \mathbf{0}$ and $y_0 \doteq 0$ for convenience.

The use of a Student-$t$ instead of a Normal distribution is quite popular in standard single-regime GARCH literature. For regime-switching models, a Student-$t$ distribution might be seen as superfluous since the switching regime can account for large unconditional kurtosis in the data [see, *e.g.*, Haas *et al.* 2004]. However, as empirically observed by Klaassen [2002], allowing for Student-$t$ innovations within regimes can enhance the stability of the states and allows to focus on the conditional variance's behavior instead of capturing some outliers. Moreover, the Student-$t$ distribution includes the Normal distribution as the limiting case where the degrees of freedom parameter goes to infinity. We have therefore an additional flexibility in the modeling and can impose Normality by constraining the lower boundary for the degrees of freedom parameter through the prior distribution.

As pointed out in **Sect. 5.1**, the Student-$t$ specification (7.3) needs to be re-expressed to perform a convenient Bayesian estimation. This is achieved as follows:

$$\begin{aligned} y_t &= \varepsilon_t(\varpi_t \varrho h_t)^{1/2} \quad \text{for } t = 1, \dots, T \\ \varepsilon_t &\overset{iid}{\sim} \mathcal{N}(0, 1) \\ \varpi_t &\overset{iid}{\sim} \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \end{aligned}$$

where $\mathcal{N}(0,1)$ is the standard Normal and $\mathcal{IG}$ denotes the Inverted Gamma density. The degrees of freedom parameter $\nu$ characterizes the density of $\varpi_t$ as follows:

$$p(\varpi_t \mid \nu) = \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \left[\Gamma\left(\frac{\nu}{2}\right)\right]^{-1} \varpi_t^{-\frac{\nu}{2}-1} \exp\left[-\frac{\nu}{2\varpi_t}\right] . \tag{7.4}$$

For a parsimonious expression of the likelihood function, we define the $T \times 1$ vectors $\mathbf{y} \doteq (y_1 \cdots y_T)'$, $\boldsymbol{\varpi} \doteq (\varpi_1 \cdots \varpi_T)'$ as well as $\mathbf{s} \doteq (s_1 \cdots s_T)'$ and regroup the ARCH parameters into the $3K \times 1$ vector $\boldsymbol{\alpha} \doteq (\boldsymbol{\alpha}_0' \ \boldsymbol{\alpha}_1' \ \boldsymbol{\alpha}_2')'$. The model parameters are then regrouped into the augmented set of parameters $\Theta \doteq (\psi, \boldsymbol{\varpi}, \mathbf{s})$ where $\psi \doteq (\boldsymbol{\alpha}, \boldsymbol{\beta}, \nu, P)$. Finally, we define the $T \times T$ diagonal matrix:

$$\Sigma \doteq \Sigma(\Theta) = \operatorname{diag}\big(\{\varpi_t \varrho\, \mathbf{e}_t' \mathbf{h}_t\}_{t=1}^T\big)$$

where we recall that $\varrho$, $\mathbf{e}_t$ and $\mathbf{h}_t$ are both functions of the model parameters, respectively given by:

$$\begin{aligned} \varrho(\nu) &\doteq \frac{\nu - 2}{\nu} \\ \mathbf{e}_t(s_t) &\doteq \big(\mathbb{I}_{\{s_t=1\}} \cdots \mathbb{I}_{\{s_t=K\}}\big)' \end{aligned}$$

and:

$$\mathbf{h}_t(\boldsymbol{\alpha}, \boldsymbol{\beta}) \doteq \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1} < 0\}}) y_{t-1}^2 + \boldsymbol{\beta} \odot \mathbf{h}_{t-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) .$$

We can now express the likelihood function of $\Theta$ as follows:

$$\mathcal{L}(\Theta \mid \mathbf{y}) \propto (\det \Sigma)^{-1/2} \exp\left[-\tfrac{1}{2} \mathbf{y}' \Sigma^{-1} \mathbf{y}\right] . \tag{7.5}$$

In the Bayesian approach, the vector of hidden states is considered as a parameter as implied by expression (7.5).

The likelihood function (7.5) is invariant with respect to relabeling the states (*i.e.*, the labeling of the states can be interchanged without affecting the likelihood value), which leads to a lack of identification of the state-specific parameters. So, without a prior inequality restriction on some state-specific parameters, a multimodal posterior is obtained and is difficult to interpret and summarize. To overcome this problem, we make use of the permutation sampler of Frühwirth-Schnatter [2001b] to find suitable identification constraints. The permutation sampler requires priors that are labeling invariant. Furthermore, we cannot be completely non-informative about the state specific parameters since, from a theoretical viewpoint, this would result in improper posteriors [see

Diebolt and Robert 1994]. These points have therefore to be taken into account when choosing the prior densities.

Conditionally on the $K \times K$ transition probabilities matrix $P \doteq [P_{ij}]$ where:

$$P_{ij} \doteq \mathbb{P}(s_{t+1} = j \mid s_t = i)$$

the prior on vector $\mathbf{s}$ is Markov:

$$p(\mathbf{s} \mid P) = \pi(s_1) \prod_{i=1}^{K} \prod_{j=1}^{K} P_{ij}^{N_{ij}}$$

where $N_{ij} \doteq \#\{s_{t+1} = j \mid s_t = i\}$ is the number of one-step transitions from state $i$ to $j$ in the $T \times 1$ vector $\mathbf{s}$. The mass function for the initial state, $\pi(s_1)$, is obtained by calculating the ergodic probabilities of the Markov chain. The vector of ergodic probabilities can be obtained as the sum of the columns of matrix $(A'A)^{-1}$, where the matrix $A$ is defined as follows:

$$A \doteq \begin{pmatrix} I_K - P' \\ \boldsymbol{\iota}_K' \end{pmatrix}$$

where $I_K$ is a $K \times K$ identity matrix and $\boldsymbol{\iota}_K$ a $K \times 1$ vector of ones [see Hamilton 1994, Sect.22.2].

The prior density for the $K \times K$ transition matrix $P$ is obtained by assuming that the $K$ rows are independent and that the density of the $i$th row is Dirichlet with parameter $\boldsymbol{\eta}_i \doteq (\eta_{i1} \cdots \eta_{iK})$:

$$\begin{aligned} p(P) &= \prod_{i=1}^{K} \mathcal{D}(\boldsymbol{\eta}_i) \\ &\propto \prod_{i=1}^{K} \prod_{j=1}^{K} P_{ij}^{\eta_{ij}-1} \, . \end{aligned}$$

Due to the labeling invariance assumption, we require that $\eta_{ii} \doteq \eta_p$ for $i = 1, \ldots, K$ and $\eta_{ij} \doteq \eta_q$ for $i, j \in \{1, \ldots, K; i \neq j\}$. A prior density with $\eta_p > \eta_q$ could be used to model the belief that the probability of persistence is bigger than the probability of transition.

For the scedastic function's parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we use truncated Normal densities:

$$\begin{aligned} p(\boldsymbol{\alpha}) &\propto \mathcal{N}_{3K}(\boldsymbol{\alpha} \mid \boldsymbol{\mu}_\alpha, \Sigma_\alpha) \mathbb{I}_{\{\boldsymbol{\alpha} > \mathbf{0}\}} \\ p(\boldsymbol{\beta}) &\propto \mathcal{N}_{K}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \mathbb{I}_{\{\boldsymbol{\beta} > \mathbf{0}\}} \end{aligned}$$

where we recall that $\boldsymbol{\mu}_\bullet$ and $\Sigma_\bullet$ are the hyperparameters, $\mathbf{0}$ is a vector of zeros of appropriate size and $\mathcal{N}_d$ is the $d$-dimensional Normal density ($d > 1$). The assumption of labeling invariance is fulfilled if we assume further that the hyperparameters are the same for all states. In particular, we set:

$$[\boldsymbol{\mu}_\alpha]_i \doteq \mu_{\alpha_0} \ , \quad [\Sigma_\alpha]_{ii} \doteq \sigma^2_{\alpha_0} \quad , \quad \left[\boldsymbol{\mu}_\beta\right]_i \doteq \mu_\beta \quad , \quad [\Sigma_\beta]_{ii} \doteq \sigma^2_\beta$$

for $i = 1, \ldots, K$, and:

$$[\boldsymbol{\mu}_\alpha]_i \doteq \mu_{\alpha_1} \quad , \quad [\Sigma_\alpha]_{ii} \doteq \sigma^2_{\alpha_1}$$

for $i = K+1, \ldots, 2K$, and:

$$[\boldsymbol{\mu}_\alpha]_i \doteq \mu_{\alpha_2} \quad , \quad [\Sigma_\alpha]_{ii} \doteq \sigma^2_{\alpha_2}$$

for $i = 2K+1, \ldots, 3K$, where $\mu_{\alpha_j}$, $\sigma^2_{\alpha_j}$ $(j = 0, 1, 2)$, and $\mu_\beta$, $\sigma^2_\beta$ are fixed hyperparameters. We note that matrices $\Sigma_\alpha$ and $\Sigma_\beta$ are diagonal in this case.

The prior density of the $T \times 1$ vector $\boldsymbol{\varpi}$ conditional on $\nu$ is found by noting that $\varpi_t$ are independent and identically distributed from (7.4), which yields:

$$p(\boldsymbol{\varpi} \mid \nu) = \left(\frac{\nu}{2}\right)^{\frac{T\nu}{2}} \left[\Gamma\left(\frac{\nu}{2}\right)\right]^{-T} \left(\prod_{t=1}^{T} \varpi_t\right)^{-\frac{\nu}{2}-1} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\frac{\nu}{\varpi_t}\right] .$$

Following Deschamps [2006], we choose a translated Exponential with parameters $\lambda > 0$ and $\delta \geqslant 2$ for the degrees of freedom parameter:

$$p(\nu) = \lambda \exp[-\lambda(\nu - \delta)]\mathbb{I}_{\{\delta < \nu < \infty\}} .$$

Finally, we form the joint prior by assuming prior independence between $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $(\boldsymbol{\varpi}, \nu)$ and $(\mathbf{s}, P)$ as follows:

$$p(\Theta) = p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\varpi} \mid \nu)p(\nu)p(\mathbf{s} \mid P)p(P)$$

and by combining the likelihood function (7.5) with the joint prior above, we obtain the posterior density via Bayes' rule:

$$p(\Theta \mid \mathbf{y}) \propto \mathcal{L}(\Theta \mid \mathbf{y})p(\Theta) . \tag{7.6}$$

## 7.2 Simulating the joint posterior

We draw an initial value:

$$\Theta^{[0]} \doteq (\boldsymbol{\alpha}^{[0]}, \boldsymbol{\beta}^{[0]}, \boldsymbol{\varpi}^{[0]}, \nu^{[0]}, \mathbf{s}^{[0]}, P^{[0]})$$

from the joint prior and we generate iteratively $J$ passes for $\Theta$. A single pass is decomposed as follows:

$$\begin{aligned}
\mathbf{s}^{[j]} &\sim p(\mathbf{s} \mid \boldsymbol{\alpha}^{[j-1]}, \boldsymbol{\beta}^{[j-1]}, \boldsymbol{\varpi}^{[j-1]}, \nu^{[j-1]}, P^{[j-1]}, \mathbf{y}) \\
P^{[j]} &\sim p(P \mid \mathbf{s}^{[j]}) \\
\boldsymbol{\alpha}^{[j]} &\sim p(\boldsymbol{\alpha} \mid \boldsymbol{\beta}^{[j-1]}, \boldsymbol{\varpi}^{[j-1]}, \nu^{[j-1]}, \mathbf{s}^{[j]}, \mathbf{y}) \\
\boldsymbol{\beta}^{[j]} &\sim p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}^{[j]}, \boldsymbol{\varpi}^{[j-1]}, \nu^{[j-1]}, \mathbf{s}^{[j]}, \mathbf{y}) \\
\boldsymbol{\varpi}^{[j]} &\sim p(\boldsymbol{\varpi} \mid \boldsymbol{\alpha}^{[j]}, \boldsymbol{\beta}^{[j]}, \nu^{[j-1]}, \mathbf{s}^{[j]}, \mathbf{y}) \\
\nu^{[j]} &\sim p(\nu \mid \boldsymbol{\varpi}^{[j]}) \, .
\end{aligned} \tag{7.7}$$

In (7.7), only $\boldsymbol{\varpi}$ and $P$ can be simulated from known expressions. Draws of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are achieved by a multivariate extension of the methodology proposed by Nakatsuma [1998, 2000]. The generation of state vector $\mathbf{s}$ is made by using the Forward Filtering Backward Sampling (henceforth FFBS) algorithm described in Chib [1996]. Finally, sampling $\nu$ is achieved by an efficient rejection technique.

As pointed out previously, the likelihood function and the joint prior are labeling invariant. Consequently, the joint posterior density in (7.6) will also be invariant and hence exhibit, at least theoretically, $K!$ different modes. Therefore, it is important to carefully select constraints to identify the model. In effect, a constraint that ignores the geometry of the posterior density will not lead to a unique labeling and can introduce a bias toward the constraint, as shown in Frühwirth-Schnatter [2001b]. If a suitable identifying restriction is not available or is not known a priori, an elegant solution to determine these constraints is to use the *random permutation sampler* proposed by Frühwirth-Schnatter [2001b]. In this version of the permutation sampler, each pass of the MCMC scheme is followed by a random permutation of the regime definitions. Formally, a random permutation $\{\Pi_1, \dots \Pi_K\}$ of $\{1, \dots, K\}$ is selected with probability $\frac{1}{K!}$. Then, for $i, j \in \{1, \dots, K\}$, the element $(i, j)$ of $P$ is replaced by the element with indices $(\Pi_i, \Pi_j)$. The hidden states process $\{s_t\}$ is substituted by $\{\Pi_{s_t}\}$. Finally, for $k = 1, \dots, K$, parameter $\alpha_0^k$ is replaced by $\alpha_0^{\Pi_k}$, parameter $\alpha_1^k$ by $\alpha_1^{\Pi_k}$, parameter $\alpha_2^k$ by $\alpha_2^{\Pi_k}$ and parameter $\beta^k$ by $\beta^{\Pi_k}$. Hence, relabeling only affects the scedastic function's parameters, the state process and the transition probabilities while the vector $\boldsymbol{\varpi}$ and the degrees of freedom parameter $\nu$ remain unchanged.

The random permutation sampler by Frühwirth-Schnatter [2001b] is used to improve the mixing of the MCMC sampler and to explore the full unconstrained parameter space. Then post-processing the MCMC output of the random per-

mutation sampler in an exploratory way, by plotting scatter plots for instance, can suggest an appropriate identification constraint, such as:

$$\beta^1 \; < \; \ldots \; < \; \beta^K \tag{7.8}$$

meaning, in this particular case, that the MS-GJR model can be identified through inequalities on the parameter $\beta$ between regimes. At this stage, the model is estimated again under the constraint (7.8) by enforcing the corresponding permutation of the regimes. This version of the permutation sampler is referred to as the *constrained permutation sampler*. At each sweep of the sampler, we test whether the constraint is fulfilled. If not, we order the pairs $\{1, \beta^1\}, \ldots, \{K, \beta^K\}$ with respect to the second component. The first component $\{\Pi_1, \ldots, \Pi_K\}$ of the ordered pairs defines the correct permutation of reordering the state parameters and this permutation is applied to the state-specific components, as this was done for the random permutation sampler. If the model is identifiable up to permutations of the states and satisfies certain regularity conditions, the constrained posterior density will exhibit a single mode. Note that the selection of the constraint (7.8) is arbitrary, because there exist $K!$ different ways of formulating constraints which render the model identified, namely:

$$\beta^{\Pi_1} \; < \; \ldots \; < \; \beta^{\Pi_K}$$

for all permutations $\{\Pi_1, \ldots, \Pi_K\}$ of $\{1, \ldots, K\}$. At this stage, if label switching still occurs, this might indicate that the inequality restriction (7.8) is not well suited or that the number $K$ of chosen regimes is too large [see Frühwirth-Schnatter 2006, Sect.4.2]. We will now present the derivation for the full conditionals appearing in the MCMC scheme (7.7).

### 7.2.1 Generating vector s

The generation of posterior samples for the $T \times 1$ vector $\mathbf{s}$ is carried out in a multi-move manner by using the FFBS algorithm. We refer the reader to Chib [1996] and Frühwirth-Schnatter [2006, Chap.11] for a detailed presentation of this procedure. We mention however that the FFBS approach can be used since the conditional density of $y_t$ only depends on the current regime which is a consequence of the definition for the conditional variance $h_t \doteq \mathbf{e}_t' \mathbf{h}_t$. Other specifications for the conditional variance in Gray [1996] or Klaassen [2002] for instance, do not allow for such an approach, as noted in Kaufmann and Frühwirth-Schnatter [2002, Sect.6.3]. The application of the FFBS algorithm has the potential advantage that the states are updated as a single block,

which avoids superfluous correlation in the vector's components, and therefore enhances the sampler's efficiency [see Frühwirth-Schnatter 2006, Sect.11.5.6].

### 7.2.2 Generating matrix $\boldsymbol{P}$

The full conditional density of the transition matrix can be derived without regard to the sampling model since $P$ becomes independent of $\Theta$ and $\mathbf{y}$ given the vector of states. Indeed, the posterior density is obtained as follows:

$$\begin{aligned} p(P \mid \mathbf{s}) &\propto p(\mathbf{s} \mid P)p(P) \\ &\propto \left( \pi(s_1) \prod_{i=1}^{K} \prod_{j=1}^{K} P_{ij}^{N_{ij}} \right) \times \left( \prod_{i=1}^{K} \prod_{j=1}^{K} P_{ij}^{\eta_{ij}-1} \right) \\ &\propto \prod_{i=1}^{K} \prod_{j=1}^{K} P_{ij}^{N_{ij}+\eta_{ij}-1} \\ &\propto \prod_{i=1}^{K} \mathcal{D}(\widehat{\boldsymbol{\eta}}_i) \end{aligned} \tag{7.9}$$

where $\widehat{\boldsymbol{\eta}}_i \doteq (N_{i1}+\eta_{i1} \cdots N_{iK}+\eta_{iK})$ and $N_{ij} \doteq \#\{s_{t+1} = j \mid s_t = i\}$ is the total number of one-step transitions from state $i$ to state $j$ in the vector $\mathbf{s}$. The rows of matrix $P$ are independent a posteriori and the $i$th row follows a Dirichlet density with parameter $\widehat{\boldsymbol{\eta}}_i$.

### 7.2.3 Generating the GJR parameters

The methodology used to draw vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be viewed as a multivariate extension of the approach proposed in **Chap. 4** for the single-regime GJR model. Let us consider the following $K \times 1$ vector:

$$\mathbf{w}_t \doteq \frac{y_t^2 \boldsymbol{\iota}_K}{\tau_t} - \mathbf{h}_t$$

where we define $\tau_t \doteq \varpi_t \rho$ for convenience and recall that $\boldsymbol{\iota}_K$ is a $K \times 1$ vector of ones. In order to simplify the notations further, we define $v_t \doteq \frac{y_t^2}{\tau_t}$ which yields $\mathbf{w}_t = v_t \boldsymbol{\iota}_K - \mathbf{h}_t$. From there, we can transform the expression for the vector of conditional variances as follows:

$$
\begin{aligned}
\mathbf{h}_t &= \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1}<0\}}) y_{t-1}^2 + \boldsymbol{\beta} \odot \mathbf{h}_{t-1} \\
\Leftrightarrow (v_t \boldsymbol{\iota}_K - \mathbf{w}_t) &= \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1}<0\}}) y_{t-1}^2 \\
&\quad + \boldsymbol{\beta} \odot (v_{t-1} \boldsymbol{\iota}_K - \mathbf{w}_{t-1}) \\
\Leftrightarrow v_t \boldsymbol{\iota}_K &= \boldsymbol{\alpha}_0 + \left[\tau_{t-1}(\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1}<0\}}) + \boldsymbol{\beta}\right] \odot v_{t-1} \boldsymbol{\iota}_K \\
&\quad - \boldsymbol{\beta} \odot \mathbf{w}_{t-1} + \mathbf{w}_t \\
\Leftrightarrow \mathbf{w}_t &= v_t \boldsymbol{\iota}_K - \boldsymbol{\alpha}_0 - \left[\tau_{t-1}(\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1}<0\}}) + \boldsymbol{\beta}\right] \odot v_{t-1} \boldsymbol{\iota}_K \\
&\quad + \boldsymbol{\beta} \odot \mathbf{w}_{t-1} \ .
\end{aligned}
$$

Moreover, let us define $w_t \doteq \mathbf{e}_t' \mathbf{w}_t$ and note that $w_t$ can be written as follows:

$$
\begin{aligned}
w_t &\doteq \mathbf{e}_t' \mathbf{w}_t \\
&= v_t - h_t = \left( \frac{y_t^2}{\varpi_t \varrho h_t} - 1 \right) h_t \\
&= \left(\chi_1^2 - 1\right) h_t
\end{aligned}
$$

where $\chi_1^2$ denotes a Chi-squared variable with one degree of freedom. This comes from the fact that the conditional distribution of $y_t$ is Normal with zero mean and variance $\varpi_t \varrho h_t$. Therefore, the conditional mean of $w_t$ is zero and the conditional variance is $2h_t^2$. As in the single-regime GJR model, this variable can be approximated by $z_t$, a Normal variable with a mean of zero and a variance of $2h_t^2$. The variable $z_t$ can be further expressed as $z_t \doteq \mathbf{e}_t' \mathbf{z}_t$ where $\mathbf{z}_t$ is a function of vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ given by:

$$
\begin{aligned}
\mathbf{z}_t(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= v_t \boldsymbol{\iota}_K - \boldsymbol{\alpha}_0 - \left[\tau_{t-1}(\boldsymbol{\alpha}_1 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \boldsymbol{\alpha}_2 \mathbb{I}_{\{y_{t-1}<0\}}) + \boldsymbol{\beta}\right] \odot v_{t-1} \boldsymbol{\iota}_K \\
&\quad + \boldsymbol{\beta} \odot \mathbf{z}_{t-1}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \ .
\end{aligned}
\tag{7.10}
$$

Then, we construct the $T \times 1$ vector $\mathbf{z} \doteq (z_1 \cdots z_T)'$ where $z_t \doteq \mathbf{e}_t' \mathbf{z}_t$ as well as the $T \times T$ diagonal matrix:

$$
\Lambda \doteq \Lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \operatorname{diag}\left(\{2\mathbf{e}_t' \mathbf{h}_t^2(\boldsymbol{\alpha}, \boldsymbol{\beta})\}_{t=1}^T\right)
$$

and express the approximate likelihood function of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as follows:

$$
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \varpi, \nu, \mathbf{s}, \mathbf{y}) \propto (\det \Lambda)^{-1/2} \exp\left[-\tfrac{1}{2} \mathbf{z}' \Lambda^{-1} \mathbf{z}\right] \ . \tag{7.11}
$$

As will be shown hereafter, the construction of the proposal densities for parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is based on this likelihood function.

**Generating vector $\boldsymbol{\alpha}$**

First, we note that the function $\mathbf{z}_t(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (7.10) can be expressed as a linear function of vector $\boldsymbol{\alpha}$. To show this, we simply extend the argument of the single-regime GJR model by using appropriate recursive transformations. More precisely, the $i$th component of the $K \times 1$ vector $\mathbf{z}_t$ can be written as follows:

$$[\mathbf{z}_t]_i = v_t - \begin{pmatrix} l_t^*(\beta^i) & v_t^*(\beta^i) & v_t^{**}(\beta^i) \end{pmatrix} \begin{pmatrix} \alpha_0^i \\ \alpha_1^i \\ \alpha_2^i \end{pmatrix}$$

with the recursive transformations $l_t^*$, $v_t^*$ and $v_t^{**}$ given by:

$$\begin{aligned} l_t^*(\beta^i) &\doteq 1 + \beta^i l_{t-1}^*(\beta^i) \\ v_t^*(\beta^i) &\doteq y_{t-1}^2 \mathbb{I}_{\{y_{t-1} \geqslant 0\}} + \beta^i v_{t-1}^*(\beta^i) \\ v_t^{**}(\beta^i) &\doteq y_{t-1}^2 \mathbb{I}_{\{y_{t-1} < 0\}} + \beta^i v_{t-1}^{**}(\beta^i) \end{aligned} \tag{7.12}$$

where $l_0^* = v_0^* = v_0^{**} \doteq 0$. We notice that $l_t^*(\bullet)$, $v_t^*(\bullet)$ and $v_t^{**}(\bullet)$ in (7.12) are similar to the recursive transformations used for the single-regime GJR model. Let us now regroup the recursive values into a $K \times 3K$ matrix $C_t$ as follows:

$$C_t \doteq \begin{pmatrix} l_t^*(\beta^1) & 0 & \cdots & 0 & v_t^*(\beta^1) & 0 & \cdots & 0 & v_t^{**}(\beta^1) & 0 & \cdots & 0 \\ 0 & l_t^*(\beta^2) & 0 & \vdots & 0 & v_t^*(\beta^2) & 0 & \vdots & 0 & v_t^{**}(\beta^2) & 0 & \vdots \\ \vdots & 0 & \ddots & 0 & \vdots & 0 & \ddots & 0 & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & l_t^*(\beta^K) & 0 & \cdots & 0 & v_t^*(\beta^K) & 0 & \cdots & 0 & v_t^{**}(\beta^K) \end{pmatrix}.$$

It is straightforward to show that $\mathbf{z}_t = v_t \boldsymbol{\iota}_K - C_t \boldsymbol{\alpha}$, and since $z_t \doteq \mathbf{e}_t' \mathbf{z}_t$ we get $z_t = v_t - \mathbf{e}_t' C_t \boldsymbol{\alpha}$. Then, by defining the $T \times 1$ vectors $\mathbf{z} \doteq (z_1 \cdots z_T)'$ and $\mathbf{v} \doteq (v_1 \cdots v_T)'$ as well as the $T \times 3K$ matrix $C$ whose $t$th row is $\mathbf{e}_t' C_t$, we end up with $\mathbf{z} = \mathbf{v} - C\boldsymbol{\alpha}$ which is the desired linear expression for $\mathbf{z}$. The proposal density to sample vector $\boldsymbol{\alpha}$ is obtained by combining the approximate likelihood (7.11) and the prior density by Bayes' update:

$$q_\alpha(\boldsymbol{\alpha} \mid \widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y}) \propto \mathcal{N}_{3K}(\boldsymbol{\alpha} \mid \widehat{\boldsymbol{\mu}}_\alpha, \widehat{\Sigma}_\alpha) \mathbb{I}_{\{\boldsymbol{\alpha} > \mathbf{0}\}} \tag{7.13}$$

with:

$$\begin{aligned} \widehat{\Sigma}_\alpha^{-1} &\doteq C' \widetilde{\Lambda}^{-1} C + \Sigma_\alpha^{-1} \\ \widehat{\boldsymbol{\mu}}_\alpha &\doteq \widehat{\Sigma}_\alpha (C' \widetilde{\Lambda}^{-1} \mathbf{v} + \Sigma_\alpha^{-1} \boldsymbol{\mu}_\alpha) \end{aligned}$$

where the $T \times T$ diagonal matrix $\widetilde{\Lambda} \doteq \text{diag}\big(\{2\mathbf{e}_t'\mathbf{h}_t^2(\widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta})\}_{t=1}^T\big)$ and $\widetilde{\boldsymbol{\alpha}}$ is the previous draw of $\boldsymbol{\alpha}$ in the M-H sampler. A candidate $\boldsymbol{\alpha}^\star$ is sampled from this proposal density and accepted with probability:

$$\min\left\{\frac{p(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}, \boldsymbol{\varpi}, \nu, \mathbf{s}, P \mid \mathbf{y})}{p(\widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{\varpi}, \nu, \mathbf{s}, P \mid \mathbf{y})} \frac{q_\alpha(\widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^\star, \boldsymbol{\beta}, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y})}{q_\alpha(\boldsymbol{\alpha}^\star \mid \widetilde{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y})}, 1\right\} .$$

**Generating vector $\boldsymbol{\beta}$**

The function $\mathbf{z}_t(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (7.10) could be expressed, in the previous section, as a linear function of $\boldsymbol{\alpha}$ but cannot be expressed as a linear function of vector $\boldsymbol{\beta}$. To overcome this problem, we linearize the $K \times 1$ vector $\mathbf{z}_t(\boldsymbol{\beta})$ by a first order Taylor expansion at point $\widetilde{\boldsymbol{\beta}}$:

$$\mathbf{z}_t(\boldsymbol{\beta}) \simeq \mathbf{z}_t(\widetilde{\boldsymbol{\beta}}) + \left.\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} \times (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})$$

where $\widetilde{\boldsymbol{\beta}}$ is the previous draw of $\boldsymbol{\beta}$ in the M-H sampler. Furthermore, let us define the following:

$$\mathbf{r}_t \doteq \mathbf{z}_t(\widetilde{\boldsymbol{\beta}}) + G_t\widetilde{\boldsymbol{\beta}} \quad , \quad G_t \doteq -\left.\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}\boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}} \tag{7.14}$$

where the $K \times K$ matrix $G_t$ can be computed by the following recursion:

$$G_t \doteq v_{t-1} I_K - Z_{t-1} + G_{t-1}\widetilde{\boldsymbol{\beta}}$$

where $Z_{t-1}$ is a $K \times K$ diagonal matrix with $\mathbf{z}_{t-1}(\widetilde{\boldsymbol{\beta}})$ in its diagonal, $I_K$ is a $K \times K$ identity matrix and $G_0$ is a $K \times K$ matrix of zeros. This recursion is simply obtained by differentiating (7.10) with respect to $\boldsymbol{\beta}$. From the definitions in (7.14) we get $\mathbf{z}_t \simeq \mathbf{r}_t - G_t\boldsymbol{\beta}$ and the approximation for $z_t$ is obtained as $z_t \simeq r_t - \mathbf{e}_t'G_t\boldsymbol{\beta}$ where $r_t \doteq \mathbf{e}_t'\mathbf{r}_t$. Let us now define the $T \times 1$ vector $\mathbf{r} \doteq (r_1 \cdots r_T)'$ as well as the $T \times K$ matrix $G$ whose $t$th row is $\mathbf{e}_t'G_t$. It turns out that $\mathbf{z} \simeq \mathbf{r} - G\boldsymbol{\beta}$, thus we can approximate the exponential of the approximate likelihood (7.11) with:

$$\exp\left[-\tfrac{1}{2}(\mathbf{r} - G\boldsymbol{\beta})'\Lambda^{-1}(\mathbf{r} - G\boldsymbol{\beta})\right] .$$

The proposal density to sample vector $\boldsymbol{\beta}$ is obtained by combining this approximation with the prior density by Bayes' update:

$$q_\beta(\boldsymbol{\beta} \mid \boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y}) \propto \mathcal{N}_K(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\mu}}_\beta, \widehat{\Sigma}_\beta)\mathbb{I}_{\{\boldsymbol{\beta}>\mathbf{0}\}} \tag{7.15}$$

with:

$$\widehat{\Sigma}_\beta^{-1} \doteq G'\widetilde{\Lambda}^{-1}G + \Sigma_\beta^{-1}$$
$$\widehat{\mu}_\beta \doteq \widehat{\Sigma}_\beta(G'\widetilde{\Lambda}^{-1}\mathbf{r} + \Sigma_\beta^{-1}\boldsymbol{\mu}_\beta)$$

where the $T \times T$ diagonal matrix $\widetilde{\Lambda} \doteq \text{diag}\big(\{2\mathbf{e}_t'\mathbf{h}_t^2(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}})\}_{t=1}^T\big)$. A candidate $\boldsymbol{\beta}^\star$ is sampled from this proposal density and accepted with probability:

$$\min\left\{\frac{p(\boldsymbol{\alpha}, \boldsymbol{\beta}^\star, \boldsymbol{\varpi}, \nu, \mathbf{s} \mid \mathbf{y})}{p(\boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\varpi}, \nu, \mathbf{s} \mid \mathbf{y})} \frac{q_\beta(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}^\star, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y})}{q_\beta(\boldsymbol{\beta}^\star \mid \boldsymbol{\alpha}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\varpi}, \nu, \mathbf{s}, \mathbf{y})}, 1\right\} .$$

### 7.2.4 Generating vector $\boldsymbol{\varpi}$

The components of $\boldsymbol{\varpi}$ are independent a posteriori and the full conditional posterior of $\varpi_t$ is obtained as follows:

$$\begin{aligned} p(\varpi_t \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \nu, \mathbf{s}, \mathbf{y}) &\propto \mathcal{L}(\Theta \mid \mathbf{y}) p(\varpi_t \mid \nu) \\ &\propto \varpi_t^{-\frac{(\nu+3)}{2}} \exp\left[-\frac{b_t}{\varpi_t}\right] \end{aligned} \tag{7.16}$$

with:

$$b_t \doteq \frac{1}{2}\left[\frac{y_t^2}{\varrho h_t} + \nu\right]$$

where we recall that $h_t \doteq \mathbf{e}_t'\mathbf{h}_t(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\varrho \doteq \frac{\nu-2}{\nu}$. Expression (7.16) is the kernel of an Inverted Gamma density with parameters $\frac{\nu+1}{2}$ and $b_t$.

### 7.2.5 Generating parameter $\boldsymbol{\nu}$

Draws from $p(\nu \mid \boldsymbol{\varpi})$ are made by optimized rejection sampling from a translated Exponential source density. This is achieved by following the lines of **Sect. 5.2.4**.

Finally, we note that the computer code and the correctness of the algorithm are tested as in previous chapters; the testing methodology is applicable to the constrained as well as unconstrained versions of the permutation sampler.

## 7.3 An application to the Swiss Market Index

We apply our Bayesian estimation method to demeaned daily log-returns $\{y_t\}$ of the Swiss Market Index (henceforth SMI). The sample period is from November 12, 1990, to December 16, 2005 for a total of 3'800 observations and the log-returns are expressed in percent. The data set is freely available from the website

`http://www.finance.yahoo.com`. Note that September 11, 2001, has not been recorded by the data provider since the stock markets closed after the terrorist attacks for a few days. From this time series, the first 2'500 observations (up to November 2001), which represent slightly less than two third of the data set, are used to estimate the model while the remaining 1'300 log-returns are used in a forecasting performance analysis.

The time series under investigation is plotted in the upper part of **Fig. 7.1** where the vertical line delimits the in- and out-of-sample observation windows. We test for autocorrelation in the times series by testing the joint nullity of autoregressive coefficients for $\{y_t\}$. We estimate the regression with autoregressive coefficients up to lag 15 and compute the covariance matrix using the White estimate. The $p$-value of the Wald test is 0.5299 which does not support the presence of autocorrelation. When testing for the autocorrelation in the series of squared observations $\{y_t^2\}$, we strongly reject the absence of autocorrelation. This is in line with the autocorrelogram of $\{y_t^2\}$ plotted in the lower part of **Fig. 7.1**. The autocorrelations are large and significantly different from zero up to lag 70. As an additional data analysis, we test for unit root using the test by Phillips and Perron [1988]. The test strongly rejects the $I(1)$ hypothesis.

We estimate the single-regime GJR$(1,1)$ model as well as the two-state Markov-switching GJR$(1,1)$ model henceforth referred to as GJR and MS-GJR for convenience. Both models are estimated using the MCMC scheme presented in **Sect. 7.2**. The estimation of the GJR model is obtained as a simplified version of the algorithm when $K = 1$ by setting the $T \times 1$ vector $\mathbf{s}$ to a vector of ones and omitting the generation of the transition matrix. For the hyperparameters on priors $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\beta})$, we set $\mu_{\alpha_i} (i = 0, 1, 2)$ and $\mu_\beta$ to zero mean vectors and choose diagonal covariance matrices for $\Sigma_{\alpha_i} (i = 0, 1, 2)$ and $\Sigma_\beta$. The variances are set to $\sigma^2_{\alpha_i} = \sigma^2_\beta = 10'000$ $(i = 0, 1, 2)$ so we do not introduce tight prior information in our estimation. In the case of the prior on the degrees of freedom parameter, we choose $\lambda = 0.01$ and $\delta = 2$; this therefore ensures the existence for the conditional variance. Finally, the hyperparameters for the prior on the transition probabilities are set to $\eta_{ii} = 2$ and $\eta_{ij} = \eta_{ji} = 1$ for $i, j \in \{1, 2\}$ so that we have a prior belief that the probabilities of persistence are bigger than the probabilities of transition.

For both models, we run two chains for 50'000 iterations each and assess the convergence of the sampler by using the diagnostic test by Gelman and Rubin [1992]. The convergence appears rather quickly, but we nevertheless consider the first half of the iterations as a burn-in phase for precaution. For the GJR model, the acceptance rates range from 88% for vector $\boldsymbol{\alpha}$ to 97% for $\beta$ indicating that the proposal densities are close to the exact conditional posteriors. The one-

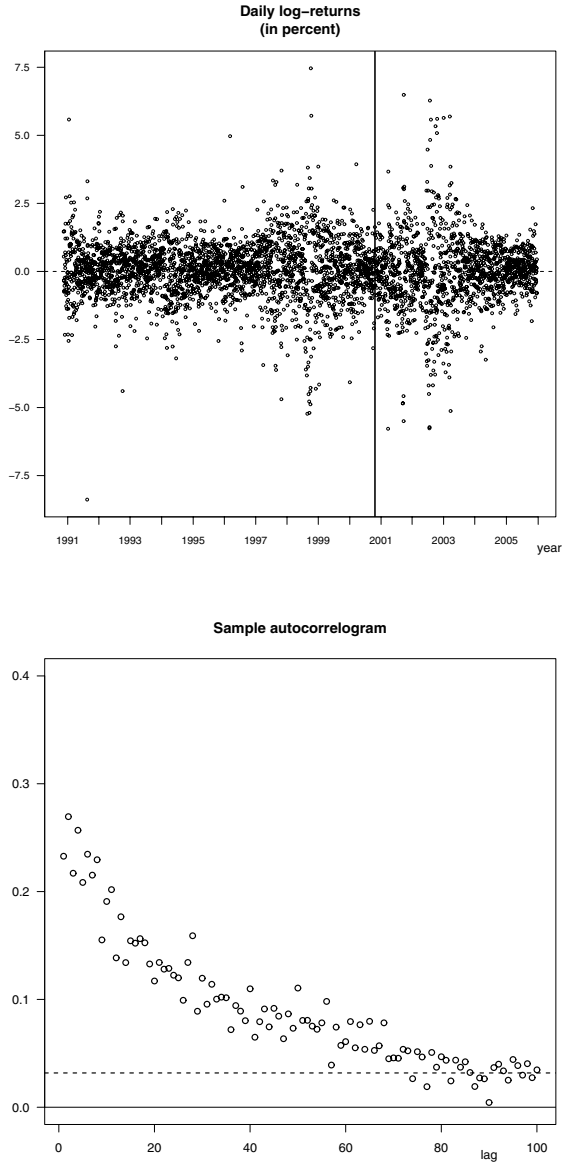


**Fig. 7.1.** SMI daily log-returns (upper graph) and sample autocorrelogram of the squared log-returns up to lag 100 (lower graph). The vertical line in the upper graph delimits the in-sample and out-of-sample observation windows.

lag autocorrelations in the chain range from 0.52 for $\alpha_1$ to 0.96 for $\beta$ which is reasonable. For the MS-GJR model, the random permutation sampler is run first to determine suitable identification constraints. In **Fig. 7.2**, we show the contour

plots of the posterior density for $(\beta^k, \alpha_0^k)$, $(\beta^k, \alpha_1^k)$ and $(\beta^k, \alpha_2^k)$, respectively. Note that the state value $k$ is arbitrary since all marginal densities contain the same information [see Frühwirth-Schnatter 2001b]. As we can notice, the bimodality of the posterior density is clear for the parameter $\beta^k$ on the three graphs, suggesting a constraint of the type $\beta^1 < \beta^2$ for identification. Therefore, the model is estimated again by imposing this constraint at each sweep in the sampler and the definition of the states is permuted if the constraint is violated. In that case, label switching only appeared 16 times after the burn-in phase thus confirming the suitability of the identification constraint. The acceptance rates obtained with the constrained version of the permutation sampler range from 22% for the vector $\boldsymbol{\alpha}$ to 93% for $\boldsymbol{\beta}$. The one-lag autocorrelations range from 0.82 for $\alpha_1^2$ to 0.97 for $\beta^2$. We keep every fifth draw from the MCMC output for both models in order to diminish the autocorrelation in the chains. The two chains are then merged to get a final sample of length 10'000. Finally, we note that a three-state Markov-switching GJR model has also been estimated. However, post-processing the MCMC output has not allowed to find a clear identification constraint.

The posterior statistics for both models are reported in **Table 7.1**. In the case of the GJR model (upper panel), we note the high persistence for the conditional variance process, measured by $\overline{\alpha} + \beta$ where $\overline{\alpha} \doteq \frac{\alpha_1+\alpha_2}{2}$, as well as the presence of the leverage effect. The estimation of the probability $\mathbb{P}(\alpha_2 > \alpha_1 \mid \mathbf{y})$ is 0.999, supporting the asymmetric behavior of the conditional variance. The low value for the estimated degrees of freedom parameter indicates conditional leptokurtosis in the data set. In the MS-GJR case (lower panel), we note also the presence of the leverage effect in both states. A comparison of the scedastic function's parameters between regimes indicates similar 95% confidence intervals for the components of the vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ while the difference for components of the $\boldsymbol{\alpha}_0$ vector is more pronounced. Indeed, for $i = 0, 1, 2$, the estimated probabilities $\mathbb{P}(\alpha_i^1 > \alpha_i^2 \mid \mathbf{y})$ are respectively 0.774, 0.397 and 0.543. As in the single-regime model, the posterior density for the degrees of freedom parameter indicates conditional leptokurtosis. We note however that the posterior mean and median are larger than for the GJR model. The posterior means for probabilities $p_{11}$ and $p_{22}$ are respectively 0.997 and 0.995 indicating infrequent mixing between states. Finally, the inefficiency factors (IF) reported in the last column of **Table 7.1** indicate that using 10'000 draws out of the MCMC sampler seems appropriate if we require that the Monte Carlo error in estimating the mean is smaller than one percent of the variation of the error due to the data. We recall that the IF are computed as the ratio of the squared numerical standard error (NSE) of the MCMC simulations and the variance estimate divided by the

number of iterations (*i.e.*, the variance of the sample mean from a hypothetical *iid* sampler). The NSE are estimated by the method of Andrews [1991], using a Parzen kernel and AR(1) pre-whitening as presented in Andrews and Monahan [1992]. As noted by Deschamps [2006], this ensures easy, optimal, and automatic bandwidth selection.

In **Fig. 7.3**, we display the marginal posterior densities for the MS-GJR model parameters. First, we note that the use of the constrained permutation sampler leads to marginal densities which are unimodal. Furthermore, we clearly notice that most of these densities are skewed. More precisely, the densities for the components of vector $\boldsymbol{\alpha}$ are right-skewed while components of $\boldsymbol{\beta}$ are left-skewed. In the case of parameters $\alpha_1^1$ and $\alpha_1^2$, the modes of the densities are close to the lower boundary of the parameter's space, suggesting that the parameters are close to zero. Finally, we can notice that the posterior densities for $p_{11}$ and $p_{22}$ are strongly left-skewed.

Some probabilistic statements on nonlinear functions of the parameters can be straightforwardly obtained by simulation from the joint posterior sample $\{\psi^{[j]}\}_{j=1}^{J}$. In particular, we can test the covariance stationarity condition and estimate the density of the unconditional variance when this condition is satisfied. Under the GJR specification, the process is covariance stationary if $\overline{\alpha} + \beta < 1$ where we recall that $\overline{\alpha} \doteq \frac{\alpha_1+\alpha_2}{2}$ for notational purposes. The estimated probability $\mathbb{P}(\overline{\alpha} + \beta < 1 \mid \mathbf{y})$ is one. Hence, the unconditional variance exists and is given by $\frac{\alpha_0}{1-\overline{\alpha}-\beta}$; the estimation of its posterior mean is 1.179 with a 95% confidence interval given by [1.173,1.189]. These estimations can be compared with the empirical variance of the process which is 1.136. In this case, the single-regime model slightly overestimates the variability of the underlying time series. For the Markov-switching model, our simulation study indicates that the process is covariance stationary in each state. The posterior mean of the unconditional variances is 0.56 in state 1 and 2.00 in state 2 with 95% confidence intervals respectively given by [0.557,0.563] and [1.992,2.012]. The unconditional variance of the process in state 1 is about four times lower than the one in state 2; we will therefore refer state 1 as the *low-volatility* regime and state 2 as the *high-volatility* regime. As found by Haas *et al.* [2004, Eq.11, p.500], the Markov-switching GARCH process is covariance stationary if $\xi(M) < 1$, where $\xi(M)$ denotes the largest eigenvalue in modulus of matrix $M$. This matrix is constructed from the model parameters and, in the case of the MS-GJR model, it is given by:

**Table 7.1.** Estimation results for the GJR model (upper panel) and MS-GJR model (lower panel).★

| GJR model | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\psi$ | $\overline{\psi}$ | $\psi_{0.5}$ | $\psi_{0.025}$ | $\psi_{0.975}$ | min | max | NSE | IF |
| $\alpha_0$ | 0.066 | 0.065 | 0.041 | 0.099 | 0.021 | 0.156 | 0.356 | 5.58 |
| $\alpha_1$ | 0.060 | 0.059 | 0.028 | 0.098 | 0.005 | 0.162 | 0.237 | 1.81 |
| $\alpha_2$ | 0.207 | 0.205 | 0.148 | 0.278 | 0.097 | 0.359 | 0.690 | 4.33 |
| $\beta$ | 0.809 | 0.809 | 0.750 | 0.861 | 0.656 | 0.911 | 1.163 | 16.22 |
| $\nu$ | 8.083 | 7.954 | 6.258 | 10.580 | 4.871 | 13.930 | 34.643 | 9.79 |
| MS-GJR model | | | | | | | | |
| $\psi$ | $\overline{\psi}$ | $\psi_{0.5}$ | $\psi_{0.025}$ | $\psi_{0.975}$ | min | max | NSE | IF |
| $\alpha_0^1$ | 0.245 | 0.241 | 0.149 | 0.362 | 0.100 | 0.518 | 2.407 | 19.26 |
| $\alpha_0^2$ | 0.184 | 0.178 | 0.089 | 0.327 | 0.046 | 0.518 | 1.939 | 10.45 |
| $\alpha_1^1$ | 0.020 | 0.015 | 0.001 | 0.063 | 0.000 | 0.145 | 0.276 | 2.61 |
| $\alpha_1^2$ | 0.027 | 0.023 | 0.001 | 0.073 | 0.000 | 0.135 | 0.302 | 2.33 |
| $\alpha_2^1$ | 0.229 | 0.224 | 0.123 | 0.361 | 0.074 | 0.534 | 1.278 | 4.21 |
| $\alpha_2^2$ | 0.220 | 0.215 | 0.136 | 0.332 | 0.090 | 0.462 | 1.140 | 5.21 |
| $\beta^1$ | 0.436 | 0.440 | 0.212 | 0.642 | 0.004 | 0.746 | 4.454 | 16.80 |
| $\beta^2$ | 0.782 | 0.785 | 0.670 | 0.866 | 0.582 | 0.907 | 2.090 | 18.33 |
| $\nu$ | 9.459 | 9.264 | 7.051 | 12.880 | 5.881 | 23.740 | 55.931 | 13.45 |
| $p_{11}$ | 0.997 | 0.997 | 0.992 | 0.999 | 0.982 | 1.000 | 0.022 | 1.23 |
| $p_{12}$ | 0.003 | 0.003 | 0.001 | 0.008 | 0.001 | 0.018 | 0.022 | 1.23 |
| $p_{21}$ | 0.005 | 0.004 | 0.001 | 0.011 | 0.001 | 0.023 | 0.027 | 1.13 |
| $p_{22}$ | 0.995 | 0.996 | 0.989 | 0.999 | 0.978 | 1.000 | 0.027 | 1.13 |

★ $\overline{\psi}$: posterior mean; $\psi_\phi$: estimated posterior quantile at probability $\phi$; min: minimum value; max: maximum value; NSE: numerical standard error ($\times 10^3$); IF: inefficiency factor (*i.e.*, ratio of the squared numerical standard error and the variance of the sample mean from a hypothetical *iid* sampler). The posterior statistics are based on 10'000 draws from the constrained posterior sample.

$$M \doteq \begin{pmatrix} p_{11}(\overline{\alpha}^1 + \beta^1) & 0 & p_{21}(\overline{\alpha}^1 + \beta^1) & 0 \\ p_{11}\alpha_1^2 & p_{11}\beta^2 & p_{21}\alpha_1^2 & p_{21}\beta^2 \\ p_{12}\beta^1 & p_{12}\alpha_1^1 & p_{22}\beta^1 & p_{22}\alpha_1^1 \\ 0 & p_{12}(\overline{\alpha}^2 + \beta^2) & 0 & p_{22}(\overline{\alpha}^2 + \beta^2) \end{pmatrix} \tag{7.17}$$

where $\overline{\alpha}^k \doteq \frac{\alpha_1^k + \alpha_2^k}{2}$. Using the posterior sample we can thus estimate the density of $\xi(M)$ by substituting the values of the draws for the model parameters in the definition (7.17). In the upper part of **Fig. 7.4**, we present the posterior density for $\xi(M)$. As we can notice, none of the values exceed one in our simulation. Thus, the model is covariance stationary. Therefore, the unconditional variance of the MS-GJR process exists and is given by:

$$h_y \doteq (\operatorname{vec} P)' \times (I_4 - M)^{-1} \times (\boldsymbol{\pi} \otimes \boldsymbol{\alpha}_0) \tag{7.18}$$

where $\boldsymbol{\pi}$ is the $2 \times 1$ vector of ergodic probabilities of the Markov chain, $I_4$ is a $4 \times 4$ identity matrix, vec denotes the vectorization operator which stacks the columns of a matrix one underneath the other and $\otimes$ denotes the Kronecker product. Derivation of formula (7.18) can be found in Haas *et al.* [2004, p.501]. The posterior density of the unconditional variance is shown in the lower part of **Fig. 7.4**. Its posterior mean is 1.134 with a 95% confidence interval of [1.128,1.139]. In this case, the confidence band for the mean contains the empirical variance of 1.136 contrary to the one in the GJR model. This suggests that the Markov-switching model is more apt to reproduce the variability of the data.

Finally, since the states vector $\mathbf{s} \doteq (s_1 \cdots s_T)'$ is considered as a parameter in the MCMC procedure, the draws $\{\mathbf{s}^{[j]}\}_{j=1}^{J}$ can also be stored and used to make inference about the smoothed probabilities. Theses probabilities are estimated as the percentage of replications of $s_t$ corresponding to regime $k$:

$$\mathbb{P}(s_t = k \mid \mathbf{y}) \approx \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}_{\{s_t^{[j]} = k\}} \, .$$

In **Fig. 7.5**, we present the smoothed probabilities for the high-volatility regime (solid line, left axis) together with the in-sample daily log-returns (circles, right axis). The 95% confidence bands are shown in dashed lines but are almost indistinguishable from the point estimates. The beginning of year 1991 is associated with the high-volatility state. Then, from the second half of 1991 to 1997, the returns are clearly associated with the low-volatility regime, with the exception of 1994. From 1997 to 2000, the model remains in the high-volatility regime with a transition during the second semester 2000 to the low-volatility state.

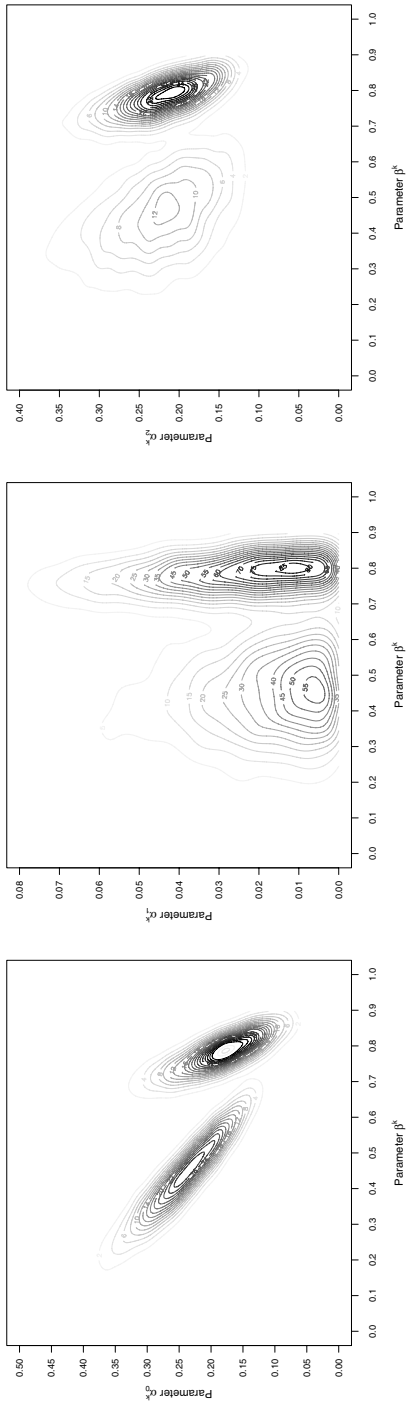


**Fig. 7.2.** Contour plots for $(\beta^k, \alpha_0^k)$, $(\beta^k, \alpha_1^k)$ and $(\beta^k, \alpha_2^k)$, respectively. The choice of $k$ is arbitrary since all marginal densities contain the same information [see Frühwirth-Schnatter 2001b]. The graphs are based on 10'000 draws from the joint posterior sample.

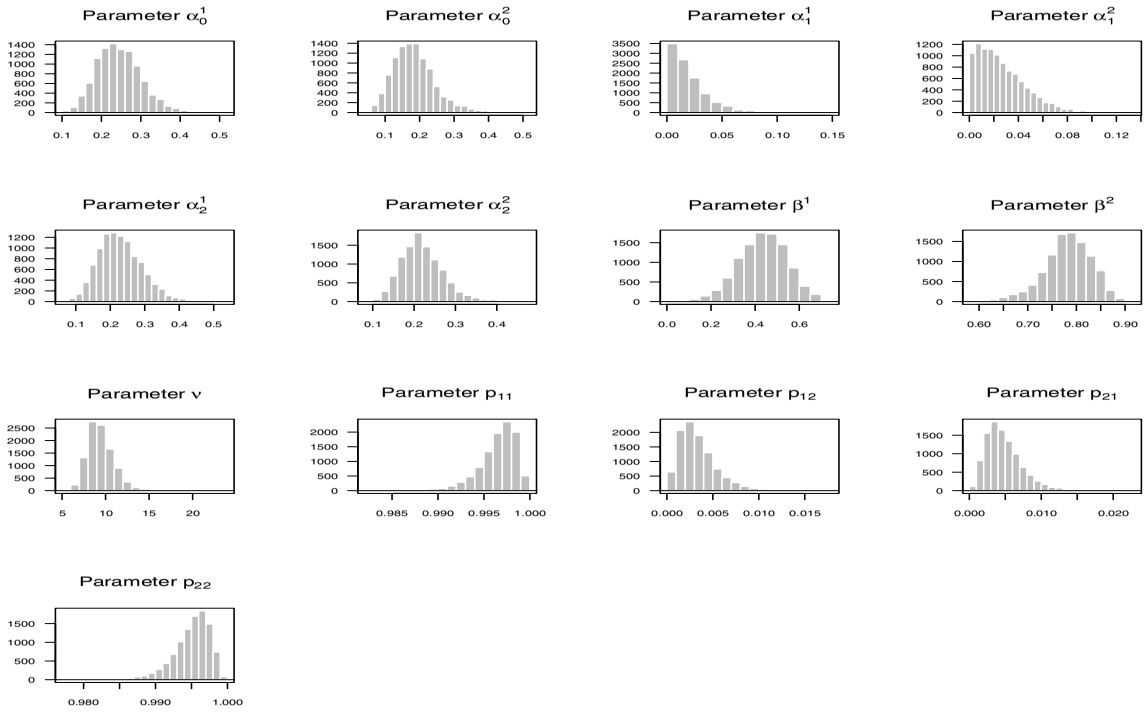


**Fig. 7.3.** Marginal posterior densities of the MS-GJR parameters. The histograms are based on 10'000 draws from the constrained posterior sample.

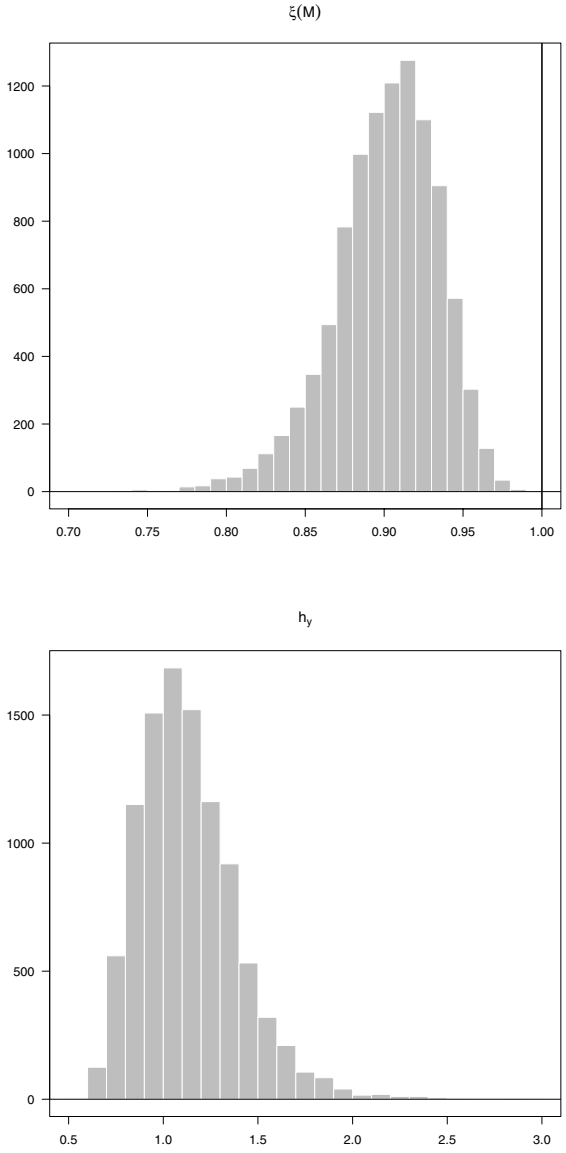


**Fig. 7.4.** Posterior densities of the covariance stationarity condition (upper graph) and the unconditional variance (lower graph) of the MS-GJR process. The histograms are based on 10'000 draws from the constrained posterior sample.

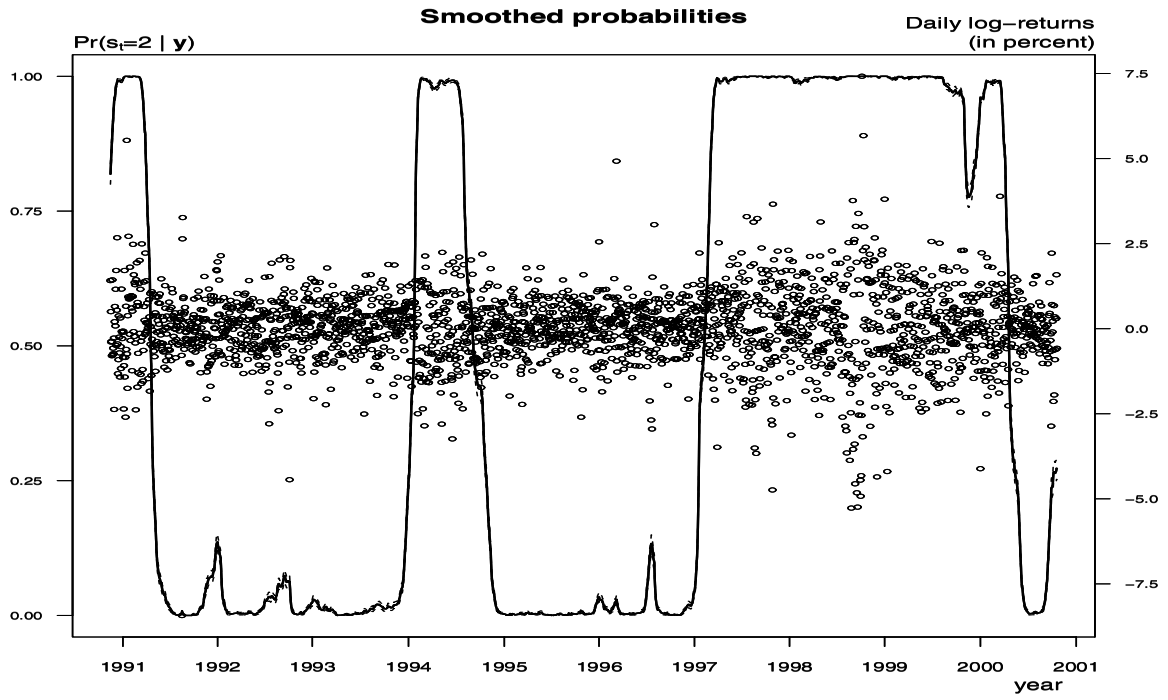


**Fig. 7.5.** Smoothed probabilities of the high-volatility state (solid line, left axis) together with the in-sample log-returns (circles, right axis). The 95% confidence bands are shown in dashed lines (they are almost indistinguishable from the point estimates).

## 7.4 In-sample performance analysis

### 7.4.1 Model diagnostics

We check for model misspecification by analyzing the predictive probabilities referred to as probability integral transforms or p-scores in the literature [see, *e.g.*, Diebold, Gunther, and Tay 1998, Kaufmann and Frühwirth-Schnatter 2002]. We make use of a simpler version of this method, as proposed by Kim, Shephard, and Chib [1998], which consists in conditioning on point estimates of $\psi$. To be meaningful, the point estimate has to be chosen when the identification is imposed. Hence, we consider the posterior mean $\overline{\psi}$ of the constrained posterior sample. Upon defining $\mathcal{F}_{t-1}$ as the information set up to time $t-1$, the (approximate) p-scores are defined as follows:

$$z_t \doteq \sum_{k=1}^{K} \mathbb{P}(Y_t \leqslant y_t \mid s_t = k, \overline{\psi}, \mathcal{F}_{t-1})\, \mathbb{P}(s_t = k \mid \overline{\psi}, \mathcal{F}_{t-1}) \; .$$

The probability $\mathbb{P}(Y_t \leqslant y_t \mid s_t = k, \overline{\psi}, \mathcal{F}_{t-1})$ can be estimated by the Student-$t$ integral and the filtered probability $\mathbb{P}(s_t = k \mid \overline{\psi}, \mathcal{F}_{t-1})$ is obtained as a byproduct from the FFBS algorithm [see Chib 1996, p.83]. Under a correct specification, the p-scores should have independent uniform distributions asymptotically [see Rosenblatt 1952]. A further transformation through the Normal integral is often applied for convenience. In this case, we consider $u_t \doteq \Phi^{-1}(z_t)$ where $\Phi^{-1}(\bullet)$ denotes the inverse cumulative standard Normal function. If the model is correct, these *generalized residuals* $\{u_t\}$ should be independent standard Normal and common tests can be used to check these features. In particular, we test the presence of autocorrelation in the series $\{u_t\}$ and $\{u_t^2\}$ using a Wald test. We also report the results of a joint test for zero mean, unit variance, zero skewness, and the absence of excess kurtosis, employing the likelihood ratio framework proposed by Berkowitz [2001]. For precisions on the testing methodology, we refer the reader to Haas *et al.* [2004, p.516].

In the case of the GJR model, the Wald statistic for testing the joint nullity of autoregressive coefficients, up to lag 15, for $u_t$ has a $p$-value of 0.0868 and for $u_t^2$, a $p$-value of 0.399. In the case of the MS-GJR model, the $p$-values are 0.0745 and 0.464, respectively. Therefore, both models seem adequate in removing the volatility clustering present in the data set. The likelihood ratio framework for testing the first four moments of the transformed residuals yields $p$-values of 0.125 for the GJR model and 0.0635 for the MS-GJR model. Overall, these results indicate no evidence of misspecification at the 5% significance level for both models.

### 7.4.2 Deviance information criterion

In order to evaluate the goodness-of-fit of the models, we use first the Deviance information criterion (henceforth DIC) introduced by Spiegelhalter *et al.* [2002]. The DIC is not intended for identification of the correct model, but rather merely as a method of comparing a collection of alternative formulations (all of which may be incorrect) and determining the most appropriate. This criterion follows from an extension of the *Deviance* proposed by Dempster [1997]. A recent article by Berg, Meyer, and Yu [2004] has illustrated the potential advantages of this information criterion in determining the appropriate stochastic volatility model. This criterion presents an interesting alternative to the Bayes factor which is often difficult to calculate, especially for models that involve many random effects, large number of unknowns or improper priors.

Let us denote the model parameters by $\theta$ for the moment. Based on the posterior density of the Deviance $D(\theta) \doteq -2 \ln \mathcal{L}(\theta \mid \mathbf{y})$ where $\mathcal{L}(\theta \mid \mathbf{y})$ is the likelihood function, the DIC consists of two terms: a component that measures the goodness-of-fit and a penalty term for any increase in model complexity. The measure of fit is obtained by taking the posterior expectation of the Deviance:

$$\overline{D} \doteq \mathbb{E}_{\theta|\mathbf{y}}\big[D(\theta)\big] \tag{7.19}$$

where $E_{\theta|\mathbf{y}}[\bullet]$ denotes the expectation with respect to the joint posterior $p(\theta \mid \mathbf{y})$. Provided that $D(\theta)$ is available in closed form, $\overline{D}$ can easily be approximated using the posterior sample by estimating the sample mean of the simulated values of $D(\theta)$. The second component measures the complexity of the model using the *effective number of parameters*, denoted by $p_D$, and is defined as the difference between the posterior mean of the Deviance and the Deviance evaluated at a point estimate $\widetilde{\theta}$:

$$p_D \doteq \overline{D} - D(\widetilde{\theta}) \; . \tag{7.20}$$

A natural candidate for $\widetilde{\theta}$ is the posterior mean $\mathbb{E}_{\theta|\mathbf{y}}(\theta)$, as suggested by Spiegelhalter *et al.* [2002]. When the density is log-concave, this point estimate ensures a positive $p_D$ due to Jensen's inequality. The DIC is then simply defined as $\text{DIC} \doteq \overline{D} + p_D$ and, given a set of models, the one with the smallest DIC has the best balance between goodness-of-fit and model complexity.

As noted in Celeux, Forbes, Robert, and Titterington [2006], the definition $\widetilde{\theta} \doteq \mathbb{E}_{\theta|\mathbf{y}}(\theta)$ is not appropriate in mixture models when no identification is imposed. Furthermore, when the state variable is discrete and considered as a parameter in $\theta$, the posterior expectation usually fails to take one of the discrete values. To overcome these difficulties, we integrate out the state vector by

considering the *observed likelihood* instead [see Celeux *et al.* 2006, Sect.3.1] and make use of the constrained posterior sample in the estimation. In the context of MS-GARCH models, the observed likelihood, also referred to as the *marginal likelihood* in Kaufmann and Frühwirth-Schnatter [2002, p.457] is obtained as follows:

$$\mathcal{L}(\psi \mid \mathbf{y}) = \prod_{t=1}^{T} \left[ \sum_{k=1}^{K} p(y_t \mid \psi, s_t = k, \mathcal{F}_{t-1}) \, \mathbb{P}(s_t = k \mid \psi, \mathcal{F}_{t-1}) \right] \tag{7.21}$$

where $p(y_t \mid \psi, s_t = k, \mathcal{F}_{t-1})$ can be estimated by the Student-$t$ density and the filtered probability $\mathbb{P}(s_t = k \mid \psi, \mathcal{F}_{t-1})$ is obtained as a byproduct from the FFBS algorithm [see Chib 1996, p.83]. The DIC is then defined as the sum of components (7.19) and (7.20), which yields:

$$\mathrm{DIC} \doteq 2 \Big\{ \ln \mathcal{L}\big(\overline{\psi} \mid \mathbf{y}\big) - 2\mathbb{E}_{\psi|\mathbf{y}} \big[ \ln \mathcal{L}(\psi \mid \mathbf{y}) \big] \Big\}$$

where we recall that $\overline{\psi} \doteq \mathbb{E}_{\psi|\mathbf{y}}(\psi)$ with $\psi \doteq (\boldsymbol{\alpha}, \boldsymbol{\beta}, \nu, P)$.

In order to make statements about the goodness-of-fit of one model relative to another, it is important to consider the uncertainty in the DIC. While the confidence interval for $\overline{D}$ can be easily obtained from the MCMC output by using spectral methods as this is done for the posterior mean, the task is more tedious in the case of $p_D$ and hence for the DIC itself. Approximation methods have been experimented in Zhu and Carlin [2000] but the *brute force* approach is still the most accurate. In this method, the variability of the DIC is estimated by running several MCMC chains and calculating the DIC's variance from the different runs. Obviously, this is extremely costly. A simpler alternative consists in running few MCMC runs and reporting the minimum and maximum DIC obtained. This gives however a crude idea of DIC's variability. In what follows, we make use of a methodology which estimates the whole distribution for the DIC based on a resampling technique. More precisely, from the joint posterior sample $\{\psi^{[j]}\}_{j=1}^{J}$, we generate randomly $B$ new posterior samples of size $J$ by using the block bootstrap technique and estimate DIC's components for these samples. By comparing the 95% confidence interval of the different DICs, we can find statistical evidence of differences in the fitting quality. With this methodology, the MCMC procedure does not need to be re-run which strongly diminishes the computing time. The choice of the block length is an important issue in the block bootstrap technique. For the block bootstrap to be effective, the length should be large enough so that it includes most of the dependence structure, but not too large so that the number of blocks becomes insufficient. In our analysis, we use the stationary bootstrap of Politis and Romano [1994]

and select the block length following the algorithm based on the spectral density estimation, as proposed by Politis and White [2004]. We apply the block length selection algorithm to each parameter's output. The maximum value is then defined as the optimal block length used for block bootstrapping the constrained posterior sample. This ad-hoc procedure allows to keep the autocorrelation in the chains as well as the cross-dependence structure between the parameters.

Results for the DIC and its components are reported in **Table 7.2**. They are based on 10'000 draws from the constrained posterior distribution. In squared brackets we give the 95% confidence interval obtained by the resampling technique using $B = 100$ replications. We keep every tenth draw from the joint posterior sample in the resampling technique in order to speed up the calculations and diminish the autocorrelation in the chains. For comparison purposes, we also consider the Bayesian information criterion introduced by Schwarz [1978] which is defined as follows:

$$\mathrm{BIC}(\psi) \doteq 2 \ln \mathcal{L}(\psi \mid \mathbf{y}) - n \ln T$$

where $n$ is the number of parameters and $T$ the number of observations. In our context, $T = 2'500$, $n = 5$ for the GJR model and $n = 11$ for the MS-GJR model (since parameters $p_{12}$ and $p_{21}$ are redundant due to the summability constraint). This criterion promotes model parsimony by penalizing models with increased model complexity (larger $n$) and sample size $T$. Hence, a model with the largest BIC is preferred. The computation of the Bayesian information criterion is based on the posterior mean $\mathbb{E}_{\psi|\mathbf{y}}\big[\mathrm{BIC}(\psi)\big]$ obtained over the 10'000 draws of the constrained posterior sample.

**Table 7.2.** Results of the DIC and BIC criteria.$^\star$

| Model | DIC | $\overline{D}$ | $p_D$ | $\mathbb{E}_{\psi\|\mathbf{y}}$(BIC) |
|---|---|---|---|---|
| GJR | 6770.4 | 6765.6 | 4.76 | -6806.07 |
| | [6769.9,6770.8] | [6765.3,6765.8] | [4.49,4.93] | (7.12) |
| MS-GJR | 6713.3 | 6704.4 | 8.84 | -6804.73 |
| | [6712.6,6713.8] | [6793.9,6794.9] | [8.49,9.04] | (12.55) |

$^\star$ DIC: Deviance information criterion; $\overline{D}$: Deviance evaluated at the posterior mean $\overline{\psi}$ (see **Table 7.1**, p.127); $p_D$: effective number of parameters; $\mathbb{E}_{\psi|\mathbf{y}}$(BIC): posterior mean of BIC($\psi$) obtained over the 10'000 draws of the constrained posterior sample; [•]: 95% confidence interval based on $B = 100$ replications of the constrained posterior sample; (•): numerical standard error ($\times 10^2$).

From **Table 7.2**, we can notice that both DIC and BIC criteria favor the MS-GJR model. Indeed, the DIC estimates based on the initial joint posterior sample is 6770.4 for the GJR model and 6713.3 for the MS-GJR model. Both 95% confidence intervals do not overlap which suggests significant improvement of the Markov-switching model. In the case of BIC, the differences between the criterion's values are less pronounced but still the Markov-switching model is favored compared to the single-regime model. If we consider now the estimations of $p_D$, we note that the estimated value is somewhat lower than five in the GJR model while about nine in the MS-GJR case. Hence, in the single-regime model, every parameter seems to be effective (or informative) when fitting the model to the data set. In the Markov-switching model however, about two third of the 13 parameters are effective. This is in line with the estimation results where it was shown that parameters $\alpha_1$ and $\alpha_2$ are almost identical across regimes. Furthermore, the $2 \times 2$ transition matrix only contains two free parameters due to the summability constraint. This suggests that the nine effective parameters of the MS-GJR model are $\alpha_0^1$, $\alpha_0^2$, $\alpha_1$, $\alpha_2$, $\beta^1$, $\beta^2$, $\nu$, $p_{11}$ and $p_{22}$.

Finally, we point out that we have also considered the posterior mode:

$$\widetilde{\psi} \doteq \arg\max_{\psi} \mathcal{L}(\psi \mid \mathbf{y})$$

in the definition of $p_D$, as suggested by Celeux *et al.* [2006, Sect.3.1]. It is argued that such a point estimate is more relevant since it depends on the posterior distribution of the whole parameter $\psi$, rather than on the marginal posterior distributions of its elements. The values of $p_D$ obtained with this new definition are larger for both models with 95% confidence intervals respectively given by [5.17,5.66] and [10.06,11.12] for the single-regime and Markov-switching models. While the preferred model remains the MS-GJR, the interpretation of parameter $p_D$ is now questionable in the GJR case since the value of $p_D$ exceeds the total number of parameters.

### 7.4.3 Model likelihood

As a second criterion to discriminate between the models under study, we consider the *model likelihood* which may be expressed as follows:

$$p(\mathbf{y}) = \int \mathcal{L}(\psi \mid \mathbf{y}) p(\psi) \mathrm{d}\psi$$

where $\mathcal{L}(\psi \mid \mathbf{y})$ is the marginal likelihood given in (7.21) and $p(\psi)$ is the joint prior density on $\psi \doteq (\boldsymbol{\alpha}, \boldsymbol{\beta}, \nu, P)$. It is clear that the model likelihood is equal to the normalizing constant of the posterior density:

$$p(\psi \mid \mathbf{y}) = \frac{\mathcal{L}(\psi \mid \mathbf{y})p(\psi)}{p(\mathbf{y})} \ .$$

The estimation of $p(\mathbf{y})$ requires the integration over the whole set of parameters $\psi$, which is a difficult task in practice, especially for complex statistical models such as ours. A full investigation of the various approaches available to estimate the model likelihood for finite mixture models can be found in Frühwirth-Schnatter [2004]. In particular, the author documents that the *bridge sampling* technique using the MCMC output of the random permutation sampler and an *iid* sample from an *importance density* $q(\psi)$ which approximates the unconstrained posterior yields the best estimator of the model likelihood (*i.e.*, the estimator with the lowest variance). Moreover, the variance of the bridge sampling estimator depends on a ratio that is bounded regardless of the tail behaviour of the importance density. This renders the estimator robust and gives more flexibility in the choice of the importance density.

First, let us recall that the bridge sampling technique of Meng and Wong [1996] is based on the following result:

$$1 = \frac{\int a(\psi)p(\psi \mid \mathbf{y})q(\psi)\mathrm{d}\psi}{\int a(\psi)q(\psi)p(\psi \mid \mathbf{y})\mathrm{d}\psi} = \frac{\mathbb{E}_q\big[a(\psi)p(\psi \mid \mathbf{y})\big]}{\mathbb{E}_{\psi|\mathbf{y}}\big[a(\psi)q(\psi)\big]} \tag{7.22}$$

where $a(\psi)$ is an arbitrary function such that $\int a(\psi)p(\psi \mid \mathbf{y})q(\psi)\mathrm{d}\psi > 0$ and $\mathbb{E}_q$ denotes the expectation with respect to the importance density $q(\psi)$. Replacing $p(\psi \mid \mathbf{y})$ by $\frac{\mathcal{L}(\psi|\mathbf{y})p(\psi)}{p(\mathbf{y})}$ in expression (7.22) yields the key identity for bridge sampling:

$$p(\mathbf{y}) = \frac{\mathbb{E}_q\big[a(\psi)\mathcal{L}(\psi \mid \mathbf{y})p(\psi)\big]}{\mathbb{E}_{\psi|\mathbf{y}}\big[a(\psi)q(\psi)\big]} \ .$$

We can estimate the model likelihood for a given function $a(\psi)$ by replacing the expectations on the right-hand side of the latter expression by sample averages. More precisely, we use MCMC draws $\{\psi^{[m]}\}_{m=1}^{M}$ from the joint posterior $p(\psi \mid \mathbf{y})$ and *iid* draws $\{\psi^{[l]}\}_{l=1}^{L}$ from the importance sampling density $q(\psi)$ to get the following approximation:

$$p(\mathbf{y}) \approx \frac{\frac{1}{L}\sum_{l=1}^{L} a(\psi^{[l]})\mathcal{L}(\psi^{[l]} \mid \mathbf{y})p(\psi^{[l]})}{\frac{1}{M}\sum_{m=1}^{M} a(\psi^{[m]})q(\psi^{[m]})} \ . \tag{7.23}$$

Meng and Wong [1996] discuss an asymptotically optimal choice for $a(\psi)$, which minimizes the expected relative error of the $p(\mathbf{y})$ estimator for *iid* draws from $p(\psi \mid \mathbf{y})$ and $q(\psi)$. This function is given by:

$$a(\psi) \propto \frac{1}{Lq(\psi) + Mp(\psi \mid \mathbf{y})} \,.$$

This special case of bridge sampling estimator is referred to as the *optimal bridge sampling* estimator by Frühwirth-Schnatter [2001a] and will be used in what follows. As the optimal choice depends on the normalized posterior $p(\psi \mid \mathbf{y})$, Meng and Wong [1996] use an iterative procedure to estimate $p(\mathbf{y})$ as a limit of a sequence $\{p_t(\mathbf{y})\}$. Based on an estimate $p_{t-1}(\mathbf{y})$ of the normalizing constant, the posterior is normalized as follows:

$$p_{t-1}(\psi \mid \mathbf{y}) \doteq \frac{\mathcal{L}(\psi \mid \mathbf{y})p(\psi)}{p_{t-1}(\mathbf{y})}$$

and a new estimate $p_t(\mathbf{y})$ is computed using approximation (7.23). This leads to the following recursion:

$$p_t(\mathbf{y}) \doteq p_{t-1}(\mathbf{y}) \times \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{p_{t-1}(\psi^{[l]}|\mathbf{y})}{Lq(\psi^{[l]})+Mp_{t-1}(\psi^{[l]}|\mathbf{y})}}{\frac{1}{M}\sum_{m=1}^{M} \frac{q(\psi^{[m]})}{Lq(\psi^{[m]})+Mp_{t-1}(\psi^{[m]}|\mathbf{y})}}$$

which can be initialized, *e.g.*, with the *reciprocal importance sampling* estimator of Gelfand and Dey [1994] given by:

$$p_0(\mathbf{y}) = \left[\frac{1}{M}\sum_{m=1}^{M} \frac{q(\psi^{[m]})}{\mathcal{L}(\psi^{[m]} \mid \mathbf{y})p(\psi^{[m]})}\right]^{-1} \,.$$

Note that this latter estimator is only based on MCMC draws from the joint posterior. Convergence of the bridge sampling technique is typically very fast in practice. In our case, the estimates converged after 3–4 iterations.

The remaining task consists in choosing an appropriate importance density to apply the bridge sampling technique. To that aim, we follow Kaufmann and Frühwirth-Schnatter [2002, pp.438–439] and Kaufmann and Scheicher [2006, pp.9–10]. The importance density is constructed in an unsupervised manner from the MCMC output of the random permutation sampler using a mixture of the proposal and conditional densities. Its construction is fully automatic and is easily incorporated in the MCMC sampler [see Frühwirth-Schnatter 2001a, p.39]. Formally, the importance density is defined as follows:

$$\begin{aligned} q(\psi) \doteq \Bigg[\frac{1}{R}\sum_{r=1}^{R} & q_\alpha(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{[r]}, \boldsymbol{\beta}^{[r]}, \boldsymbol{\varpi}^{[r]}, \nu^{[r]}, \mathbf{s}^{[r]}, \mathbf{y}) \\ & \times q_\beta(\boldsymbol{\beta} \mid \boldsymbol{\alpha}^{[r]}, \boldsymbol{\beta}^{[r]}, \boldsymbol{\varpi}^{[r]}, \nu^{[r]}, \mathbf{s}^{[r]}, \mathbf{y}) \times p(P \mid \mathbf{s}^{[r]})\Bigg] \times q_\nu(\nu) \end{aligned} \tag{7.24}$$

where:

$$\boldsymbol{\alpha}^{[r]} \quad , \quad \boldsymbol{\beta}^{[r]} \quad , \quad \boldsymbol{\varpi}^{[r]} \quad , \quad \nu^{[r]} \quad , \quad \mathbf{s}^{[r]} \quad \text{for } r = 1, \ldots, R$$

are draws from the unconstrained posterior sample, $q_\alpha(\boldsymbol{\alpha} \mid \bullet)$ is the proposal density for parameter $\boldsymbol{\alpha}$ given in (7.13), $q_\beta(\boldsymbol{\beta} \mid \bullet)$ is the proposal density for parameter $\boldsymbol{\beta}$ given in (7.15) (the normalizing constants are easily obtained as the proposals are truncated multivariate Normal densities), $p(P \mid \bullet)$ is the product of Dirichlet posterior densities for the transition probabilities given in (7.9). For the degrees of freedom parameter $\nu$, the optimized rejection technique of **Sect. 7.2.5** does not lead to a known expression for the marginal posterior on $\nu$. To tackle this problem, we approximate the marginal posterior by using a truncated skewed Student-$t$ density whose parameters are estimated by Maximum Likelihood from the posterior sample $\{\nu^{[j]}\}_{j=1}^{J}$. More precisely, the approximation may be written as follows:

$$q_\nu(\nu) \propto \mathcal{SS}(\nu \mid \widehat{\mu}, \widehat{\sigma}^2, \widehat{\tau}, \widehat{\gamma}) \mathbb{I}_{\{\nu > \delta\}}$$

where:

$$\begin{aligned} \mathcal{SS}(\nu \mid \mu, \sigma^2, \tau, \gamma) \doteq & \frac{2}{\gamma + \frac{1}{\gamma}} \frac{\Gamma\left(\frac{\tau+1}{2}\right)}{\Gamma\left(\frac{\tau}{2}\right) (\pi\tau\sigma^2)^{1/2}} \\ & \times \left[1 + \frac{(\nu-\mu)^2}{\tau\sigma^2} \left\{ \frac{1}{\gamma} \mathbb{I}_{\{\nu-\mu \geqslant 0\}} + \gamma^2 \mathbb{I}_{\{-\infty < \nu-\mu\}} \right\}\right]^{-\frac{\tau+1}{2}} \end{aligned} \tag{7.25}$$

is the skewed Student-$t$ density as defined in Fernández and Steel [1998, Eq.13, p.363]. The parameters of the density defined in (7.25) are: the location parameter $\mu$, the scale factor $\sigma^2 > 0$, the degrees of freedom parameter $\tau \geqslant 1$ and the asymmetry coefficient $\gamma > 0$. For $\gamma = 1$, the density coincides with the symmetric Student-$t$ density. In cases where $\gamma > 1$, the density is right-skewed while it is left-skewed when $\gamma < 1$. Therefore, parametrization (7.25) allows for a wide range of asymmetric and heavy-tailed densities. Moreover, the normalizing constant for $q_\nu(\nu)$ is easily obtained by conventional quadrature methods.

Some comments are in order here. First, the generation of draws from the proposal densities $q_\alpha(\boldsymbol{\alpha} \mid \bullet)$ and $q_\beta(\boldsymbol{\beta} \mid \bullet)$ is achieved by the rejection technique. While we obtain good acceptance rates in our case, this method can become very inefficient if the mass of the density is close to the domain of truncation. For these cases, we would need a more sophisticated algorithm, as proposed in Philippe and Robert [2003], Robert [1995], to draw efficiently from a truncated

multivariate Normal distribution. Second, the density $q_\nu(\nu)$ is constructed in two steps. The parameters of the the skewed Student-$t$ are first estimated by ML from the MCMC output and then the density is truncated to construct $q_\nu(\nu)$. An alternative approach would be to fit directly the truncated skewed Student-$t$ density by ML. This is however not necessary in our case since the mass of the posterior on the degrees of freedom is far from the truncation domain. Finally, generating draws from $q_\nu(\nu)$ is achieved by the rejection technique. In cases where the boundary is close to the high probability mass, alternative approaches, such as the inversion technique, are required [see, *e.g.*, Geweke 1991].

As indicated previously, the parameters of the skewed Student-$t$ density are estimated by ML using the posterior sample of $\nu$. In the case of the MS-GJR model, we obtain the following ML estimates:

$$\widehat{\mu} = 9.49 \quad , \quad \widehat{\sigma}^2 = 1.50 \quad , \quad \widehat{\tau} = 16.67 \quad \text{and} \quad \widehat{\gamma} = 1.53 \; .$$

In the upper part of **Fig. 7.6**, we display the fitted truncated skewed Student-$t$ density (in dashed line) together with the density of the posterior sample for $\nu$ (in solid line) obtained through Gaussian kernel density estimates [see Silverman 1986]. We can notice that the truncated skewed Student-$t$ density approximates the marginal closely. In the lower part of the figure, we show the marginal posterior for parameter $\beta^1$ together with the importance density computed with $R = 1'000$. As the construction of the mixture (7.24) is based on averaging over proposal densities, where the state process is sampled from the unconstrained posterior with balanced label switching, the mixture importance density is multimodal. We also notice that the importance density provides a good approximation of the marginal posterior.

In **Table 7.3**, we report the natural logarithm of the model likelihoods obtained using the reciprocal sampling estimator (second column) and the bridge sampling estimator (last column) for $M = L = 1'000$ draws. From this table, we can notice that both estimators are higher for the MS-GJR model, indicating a better in-sample fit for the regime-switching specification. As an additional discrimination criterion, we compute the (transformed) Bayes factor in favor of the MS-GJR model [see Kass and Raftery 1995, Sect.3.2]. The estimated value is $2 \times \ln \mathrm{BF} = 2 \times (-3389.66 - (-3408.04)) = 36.76$, which strongly supports the in-sample evidence in favor of the regime-switching model.

A final word about the robustness of these results is in order. It is indeed recognized that the model likelihood is sensitive to the choice of the prior density. We must therefore test whether an alternative joint prior specification would have modified the conclusion of our analysis. To answer this question, we modify the hyperparameters' values and run the sampler again. This time, we consider

**Table 7.3.** Results of the model likelihood estimators.$^\star$

| Model | $\ln p_0(\mathbf{y})$ | $\ln p(\mathbf{y})$ |
|---|---|---|
| GJR | -3405.33 (2.979) | -3408.04 (2.644) |
| MS-GJR | -3386.14 (3.109) | -3389.66 (3.191) |

$^\star$ $\ln p_0(\mathbf{y})$: natural logarithm of the model likelihood estimate using reciprocal sampling; $\ln p(\mathbf{y})$: natural logarithm of the model likelihood estimate using bridge sampling; ($\bullet$) numerical standard error of the estimators ($\times 10^2$).

slightly more informative priors for the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by choosing diagonal covariance matrices whose variances are set to $\sigma^2_{\alpha_i} = \sigma^2_\beta = 1'000$ ($i = 0, 1, 2$). As an alternative prior on the degrees of freedom parameter, we choose $\lambda = 0.02$ and $\delta = 2$, which implies a prior mean of 52. Finally, the hyperparameters for the prior on the transition probabilities are set to $\eta_{ii} = 3$ and $\eta_{ij} = \eta_{ji} = 1$ for $i, j \in \{1, 2\}$. We recall that the hyperparameters of the initial joint prior were set to $\sigma^2_{\alpha_i} = \sigma^2_\beta = 10'000$, $\lambda = 0.01$, $\delta = 2$, $\eta_{ii} = 2$ and $\eta_{ij} = \eta_{ji} = 1$. In this case, the results are similar to those obtained previously. The natural logarithm of the bridge sampling estimator is -3402.11 for the GJR model and -3388.09 for the MS-GJR model, implying a (transformed) Bayes factor of 28.04. These results are in line with the conclusion of the previous section and confirm the better fit of the Markov-switching model.

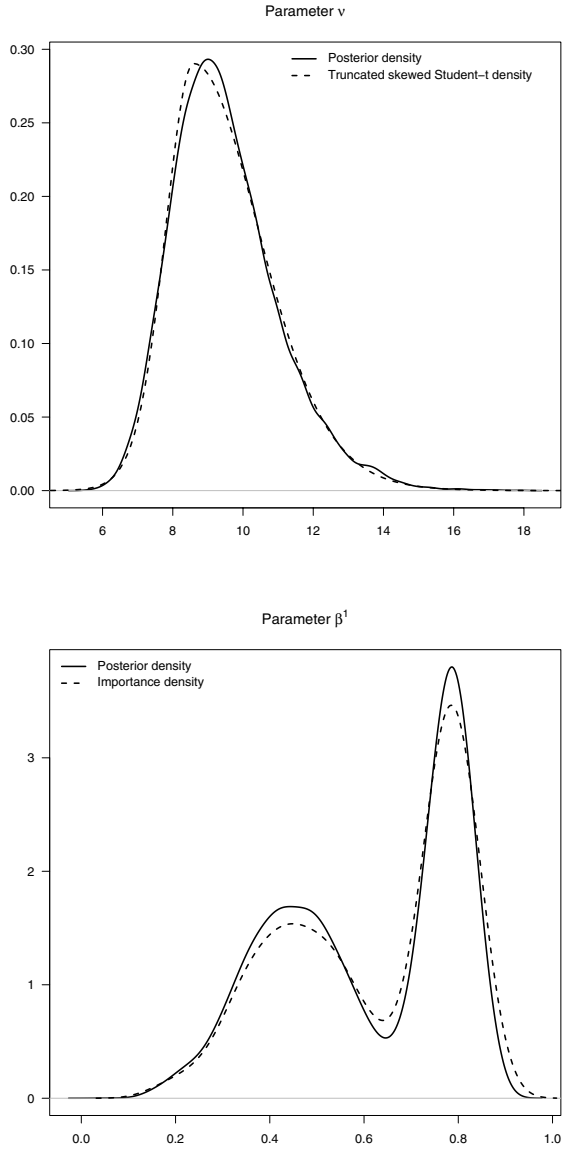


**Fig. 7.6.** Importance density (in dashed line) and marginal posterior density (in solid line) comparison. Gaussian kernel density estimates with bandwidth selected by the "Silverman's rule of thumb" criterion [see Silverman 1986, p.48]. Both graphs are based on 10'000 draws from the unconstrained posterior sample.

## 7.5 Forecasting performance analysis

In order to evaluate the ability of the competing models to predict the future behavior of the volatility process, we study the forecasted one-day ahead Value at Risk (henceforth VaR), which is a common tool to measure financial and market risks. The one-day ahead VaR at risk level $\phi \in (0, 1)$, denoted by $\text{VaR}^{\phi}$, is estimated by calculating the $\phi^c$th percentile of the one-day ahead predictive distribution, where $\phi^c \doteq (1 - \phi)$ for convenience. The predictive density is obtained by simulation from the joint posterior sample $\{\psi^{[j]}\}_{j=1}^{J}$ as follows:

$$
\begin{aligned}
s_{t+1}^{[j]} &\sim p(s_{t+1} \mid \psi^{[j]}, \mathcal{F}_t) \\
y_{t+1}^{[j]} &\sim p(y_{t+1} \mid \psi^{[j]}, s_{t+1}^{[j]}, \mathcal{F}_t)
\end{aligned}
$$

and $\text{VaR}^{\phi}$ is then simply estimated by calculating the $\phi^c$th percentile of the empirical distribution $\{y_{t+1}^{[j]}\}_{j=1}^{J}$.

In order to simulate from the predictive density over the out-of-sample observation window, the posterior sample $\{\psi^{[j]}\}_{j=1}^{J}$ should be updated using the most recent information. Consequently, forecasting the one-day ahead VaR would necessitate the estimation of the joint posterior sample at each time point in the out-of-sample observation window. However, such an approach is computationally impractical for a large data set such as ours. Combination of MCMC and importance sampling to estimate efficiently this predictive density is proposed by Gerlach, Carter, and Kohn [1999]. Nevertheless, for the sake of simplicity, we will consider the same joint posterior sample, based on the in-sample data set, when forecasting the VaR.

In addition to the static GJR and MS-GJR models, we consider a GJR model estimated on rolling windows which is the standard practice in financial risk management. This methodology relies on the assumption that older data are not available or are irrelevant due to structural breaks, which are so complicated that they cannot be modeled. We refer the reader to **Sect. 6.4.1** for a detailed presentation of this procedure. For this approach, we use 750 log-returns to estimate the model and the next 50 log-returns are used as a forecasting window. Then, the estimation and forecasting windows are moved together by 50 days ahead, so that the forecasting windows do not overlap. In this manner, the estimation methodology fulfills the recommendations of the Basel Committee in the use of internal models [see Basel Committee on Banking Supervision 1996b]. When applied to our data set, this estimation design leads to the generation of 26 estimation windows for a total of $26 \times 50 = 1'300$ out-of-sample observations. In the case of the static GJR and MS-GJR models, the first 2'500 observations of our data set are used to estimate the models while the remaining 1'300 obser-

vations are used to test their predictive performance. For the three models, the VaR predictions are obtained for the same 1'300 out-of-sample daily log-returns.

To verify the accuracy of the VaR estimates for the analyzed models, we adopt the testing methodology proposed by Christoffersen [1998]. This approach is based on the study of the random sequence $\{V_t^\phi\}$ where:

$$V_t^\phi \doteq \begin{cases} 1 & \text{if} \quad y_{t+1} < \text{VaR}_t^\phi \\ 0 & \text{else} \ . \end{cases}$$

A sequence of VaR forecasts at risk level $\phi$ has correct conditional coverage if $\{V_t^\phi\}$ is an independent and identically distributed sequence of Bernoulli random variables with parameter $\phi^c$. This hypothesis can be verified by testing jointly the independence on the series and the unconditional coverage of the VaR forecasts, *i.e.*, $\mathbb{E}(V_t^\phi) = \phi^c$, as proposed by Christoffersen [1998].

Forecasting results for the VaR are reported in **Table 7.4** for $\phi \in \{0.90, 0.95, 0.99\}$ which are typical risk levels used in financial risk management. The second and third columns give the expected and observed number of violations. The last three columns report the $p$-values for the tests of correct unconditional coverage (UC), independence (IND) and correct conditional coverage (CC). From this table, we first note that the observed number of violations for the MS-GJR model are closer to the expected values than for the static GJR model. Indeed, at the 1% significance level, the test of correct unconditional coverage is not rejected for the Markov-switching model while it is strongly rejected for the GJR model at risk level $\phi = 0.95$. The test of independence is not rejected for both models at the 1% significance level. We can notice that for risk level $\phi = 0.99$ this test is not applicable since no consecutive violations have been observed. The joint hypothesis of correct unconditional coverage and independent sequence is obtained via the test of correct conditional coverage. In the case of the MS-GJR model, $p$-values are close to 0.10 for risk levels $\phi = 0.9$ and $\phi = 0.95$ while it is 0.030 and 0.013 in the GJR case. We therefore reject the correct conditional coverage hypothesis for the static GJR model at the 5% significance level. These results indicate the better out-of-sample performance of the Markov-switching model compared to the static GJR model.

When comparing the MS-GJR model with the rolling GJR model, we can notice that both approaches perform equally well. Indeed, for both models, the test of independence is rejected at risk level $\phi = 0.90$ while the correct conditional coverage hypothesis is not rejected at the 5% significance level. Although the two models are successful in forecasting the conditional variance of the SMI log-returns, the MS-GJR model has two advantages over the rolling window

**Table 7.4.** Forecasting results of the VaR.[★]

| GJR model (static approach) | | | | | |
|---|---|---|---|---|---|
| $\phi$ | $\mathbb{E}(V_t^{\phi})$ | # | UC | IND | CC |
| 0.99 | 13 | 14 | 0.783 | NA | NA |
| 0.95 | 65 | 89 | 0.004 | 0.624 | 0.013 |
| 0.90 | 130 | 143 | 0.236 | 0.018 | 0.030 |
| GJR model (rolling windows approach) | | | | | |
| $\phi$ | $\mathbb{E}(V_t^{\phi})$ | # | UC | IND | CC |
| 0.99 | 13 | 15 | 0.586 | NA | NA |
| 0.95 | 65 | 73 | 0.318 | 0.547 | 0.506 |
| 0.90 | 130 | 126 | 0.710 | 0.032 | 0.093 |
| MS-GJR model (static approach) | | | | | |
| $\phi$ | $\mathbb{E}(V_t^{\phi})$ | # | UC | IND | CC |
| 0.99 | 13 | 13 | 1.000 | NA | NA |
| 0.95 | 65 | 80 | 0.065 | 0.323 | 0.112 |
| 0.90 | 130 | 132 | 0.854 | 0.035 | 0.107 |

[★] $\phi$: risk level; $\mathbb{E}(V_t^{\phi})$: expected number of violations; #: observed number of violations; UC: $p$-value for the correct unconditional coverage test; IND: $p$-value for the independence test; CC: $p$-value for the correct conditional coverage test; NA: not applicable.

approach. First, it is able to anticipate structural breaks in the conditional variance process. This is achieved through the estimation of the filtered probabilities $\mathbb{P}(s_t = k \mid \psi, \mathcal{F}_{t-1})$, as shown in **Fig. 7.7**. On the contrary, the rolling window methodology is merely an ad-hoc approach which is unable to forecast structural breaks. The updating frequency as well as the length of the rolling window are subjective quantities, albeit some ranges are recommended by regulators, so that different choices might lead to significant differences in the model's performance. Second, the MS-GJR model needs only to be estimated once. On the contrary, the parameters of the GJR model must be updated frequently to account for structural breaks in the time series and this can have practical consequences for risk management systems of financial institutions. This is a definite advantage of the regime-switching approach compared to the traditional rolling window methodology.

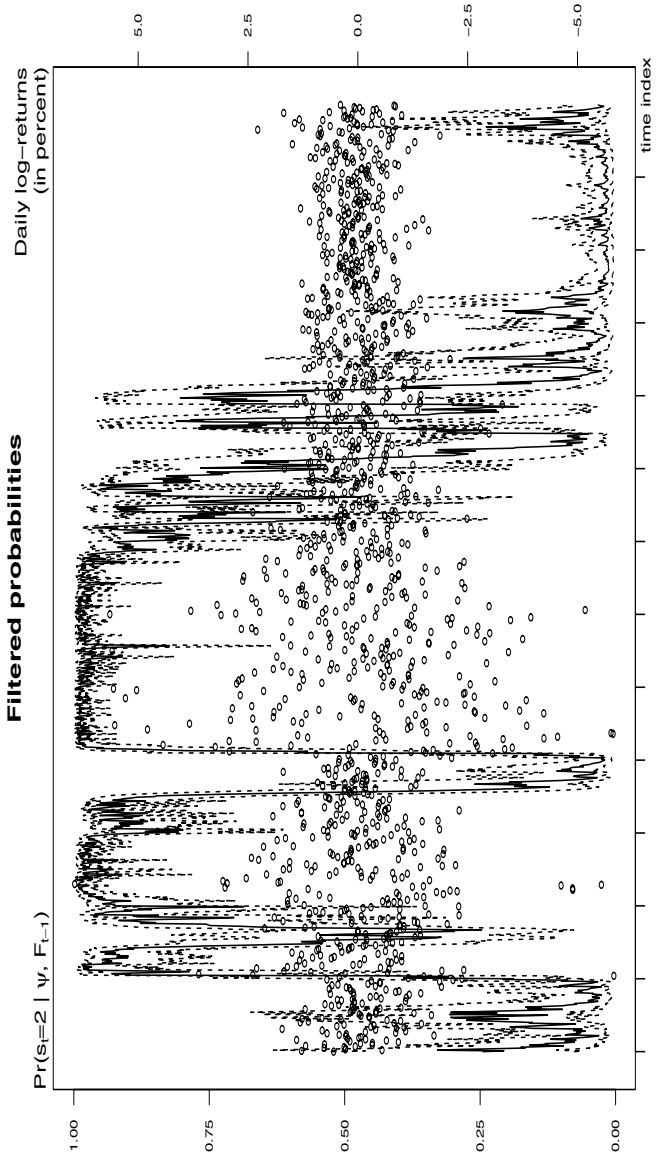


**Fig. 7.7.** Filtered probabilities of the high-volatility state (solid line, left axis) together with the out-of-sample log-returns (circles, right axis). The 95% confidence bands are shown in dashed lines.

## 7.6 One-day ahead VaR density

As emphasized in **Chap. 6**, the one-day ahead VaR risk measure can be expressed as a function of the model parameters when the underlying time series is described by a single-regime GARCH(1, 1) model. It turns out that this is also the case in the context of Markov-switching GARCH models. In effect, the one-day ahead VaR at risk level $\phi$, estimated at time $t$, can be explicitly calculated for given $\psi$ and future state $s_{t+1}$ as follows:

$$\mathrm{VaR}_t^{\phi}(\psi, s_{t+1}) \doteq \left[\varrho(\nu) \times \mathbf{e}'_{t+1}(s_{t+1})\mathbf{h}_{t+1}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right]^{1/2} \times t_{\phi^c}(\nu) \tag{7.26}$$

where we recall that $\varrho(\nu) \doteq \frac{\nu-2}{\nu}$ and $t_{\phi^c}(\nu)$ denotes the $\phi^c$th percentile of a Student-$t$ distribution with $\nu$ degrees of freedom. Hence, the VaR risk measure can be simulated from the joint posterior sample $\{\psi^{[j]}\}_{j=1}^{J}$ by first generating $s_{t+1}^{[j]}$ from the filtered probability density $p(s_{t+1} \mid \psi^{[j]}, \mathcal{F}_t)$, and then inputting the joint draw $(\psi^{[j]}, s_{t+1}^{[j]})$ in expression (7.26).

The result of this procedure is shown in **Fig. 7.8** where we plot the one-day ahead VaR density of the MS-GJR model for two distinct time points in the out-of-sample observation window. We can notice that both densities are bimodal, which is a consequence of the Markov-switching nature of the conditional variance process. At time $t = 2$'501, the VaR density gives a higher probability to larger (in absolute value) VaR values. This suggests that, at that particular point in time, the probability of being in the high volatility state is higher than being in the low-volatility regime. At time $t = 3$'500, the bimodality of the density is slightly less pronounced. In this case, the VaR density puts more mass on smaller VaR values (in absolute value). This graph shows that the density of the VaR has a particular shape in the case of the MS-GJR model. In this context, it would be interesting to determine if the loss function of an agent, and therefore the location of his optimal Bayes estimate within the VaR density, would have any influence on the forecasting performance of the model.

In order to address this question, we consider different loss functions and determine the Bayes point estimates for the VaR by solving the optimization problem (6.10) of page 85. The loss functions we consider are the Linex with a parameter $a \in \{-3, 3\}$, the absolute error loss (AEL) as well as the squared error loss (SEL); the reader is referred to **Sect. 6.4.4** for further details. We recall however that the Linex function with a positive parameter could be attributed to a regulator or risk manager whose aim is to avoid systematic failure in risk measure estimation. On the contrary, a negative parameter could be attributed to a fund manager who seeks to save risk capital since it earns little or no return at all (see **Sect. 6.3.1** for details). The AEL and SEL correspond to the

perspective of an agent for whom under- and overestimation are equally serious. The SEL leads, however, to a larger penalty for larger deviations from the true value compared to the AEL function.

The VaR risk measure obtained with the different loss functions are then tested over the 1'300 out-of-sample observations. To test the adequacy of the point estimates to reproduce the true VaR, we rely on the forecasting methodology of Christoffersen [1998] as this was done in the preceding section. The results are reported in **Table 7.5** whose second column gives the observed number of violations and the third, fourth and fifth columns report the $p$-values for the tests of correct unconditional coverage (UC), independence (IND) and correct conditional coverage (CC), respectively. From this table, we note first that the observed number of violations is close to the expected value for the Linex function with parameter $a = 3$. In this case, the test of correct unconditional coverage, at the 5% significance level, is never rejected. On the contrary, the Linex function with parameters $a = -3$ leads to the rejection of the null for risk levels $\phi = 0.95$ and $\phi = 0.99$. The null hypothesis is also rejected for the AEL and SEL point estimates at risk level $\phi = 0.95$, where the estimates systematically underestimate (in absolute value) the true VaR. The joint hypothesis of correct unconditional coverage and independence is rejected at the 5% significance level for all functions, except the Linex with $a = 3$ and the SEL at risk level $\phi = 0.9$.

From what precedes, we can thus conclude that parameter uncertainty has to be taken seriously in the context of MS-GARCH models. In particular, the choice of a given point estimate within the VaR density has a significant impact on the forecasting performance of the model. A regulator (Linex $a = 3$) whose VaR point estimate are conservative, would conclude to a good performance of the model while a fund manager (Linex $a = -3$) would systematically underestimate (in absolute value) the true VaR.

**Table 7.5.** Forecasting results of the VaR point estimates for the MS-GJR model.★

| $\phi = 0.90$, $\mathbb{E}(V_t^\phi) = 130$; Loss $\mathscr{L}$ | # | UC | IND | CC |
|---|---|---|---|---|
| Linex ($a = 3$) | 130 | 1.000 | 0.018 | 0.061 |
| Linex ($a = -3$) | 140 | 0.361 | 0.011 | 0.025 |
| AEL[a] | 133 | 0.782 | 0.011 | 0.039 |
| SEL[b] | 131 | 0.926 | 0.015 | 0.053 |
| $\phi = 0.95$, $\mathbb{E}(V_t^\phi) = 65$; Loss $\mathscr{L}$ | # | UC | IND | CC |
| Linex ($a = 3$) | 71 | 0.452 | 0.270 | 0.410 |
| Linex ($a = -3$) | 87 | 0.008 | 0.171 | 0.011 |
| AEL[a] | 84 | 0.020 | 0.228 | 0.033 |
| SEL[b] | 83 | 0.028 | 0.249 | 0.046 |
| $\phi = 0.99$, $\mathbb{E}(V_t^\phi) = 13$; Loss $\mathscr{L}$ | # | UC | IND | CC |
| Linex ($a = 3$) | 11 | 0.567 | NA | NA |
| Linex ($a = -3$) | 21 | 0.041 | NA | NA |
| AEL[a] | 17 | 0.287 | NA | NA |
| SEL[b] | 14 | 0.783 | NA | NA |

★ $\phi$: risk level; $\mathbb{E}(V_t^\phi)$: expected number of violations; #: observed number of violations; UC: $p$-value for the correct uncoverage test; IND: $p$-value for the independence test; CC: $p$-value for the correct conditional coverage test; NA: not applicable.
[a] Absolute error loss function.
[b] Squared error loss function.

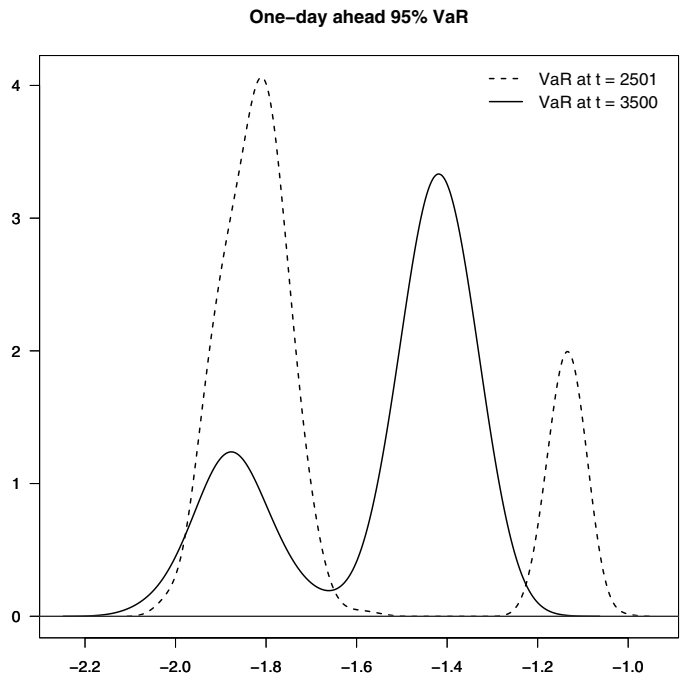


**Fig. 7.8.** Density of the one-day ahead VaR at risk level $\phi = 0.95$ for the MS-GJR model at two time points in the out-of-sample observation window. Gaussian kernel density estimates with bandwidth selected by the "Silverman's rule of thumb" criterion [see Silverman 1986, p.48]. Both graphs are based on 10'000 draws from the joint posterior density of the MS-GJR model parameters.

## 7.7 Maximum Likelihood estimation

We conclude this chapter with some comments regarding the Maximum Likelihood (henceforth ML) estimation of Markov-switching GARCH models. In this case, the estimation is handled as in Hamilton [1994, p.692], where the algorithm turns out to be a special case of the Expectation Maximization (henceforth EM) algorithm developed by Dempster, Laird, and Rubin [1977]. The classical ML approach cannot be applied directly, as the marginal likelihood where the latent process $\{s_t\}$ is integrated out, is not available in closed form. The estimation procedure is therefore decomposed into two stages. The first step consists in estimating the sequence of filtered probabilities $\{\mathbb{P}(s_t = k \mid \psi, \mathcal{F}_{t-1})\}_{t=1}^{T}$ for a fixed set of of parameters $\psi$. The second step maximizes the observed likelihood $\mathcal{L}(\psi \mid \mathbf{y})$ in expression (7.21) given this sequence of probabilities. The procedure is iterated until a given convergence criterion is satisfied. General results available for the EM algorithm indicate that the likelihood function increases in the number of iterations.

While apparently straightforward to handle, the ML estimation has practical drawbacks. Indeed, the EM algorithm guarantees a convergence to a local maximum of the likelihood, but not necessarily to the global optimum. As reported in Hamilton and Susmel [1994], many starting points are required to end up with a global maximum. Furthermore, the covariance matrix at the optimum can be extremely tedious to obtain and ad-hoc procedures are often required to get reliable results. *E.g.*, Hamilton and Susmel [1994] fix some transition probabilities to zero in order to determine the variance estimates for some model parameters. Finally, testing the null of $K$ versus $K'$ states is not possible within the ML framework since the regularity conditions for justifying the $\chi^2$ approximation of the likelihood ratio statistic do not hold.

For comparison purposes, we estimate the MS-GJR model via the ML technique. The iterative procedure described previously has been run using 20 random starting values. In all cases, the optimizer has been trapped in a local maximum or even did not converge. The convergence has only been achieved by starting the ML optimizer at the posterior mean $\overline{\psi}$ (see **Table 7.1**, p.127) obtained with the Bayesian approach.

In **Fig. 7.9**, we display the marginal densities obtained via Gaussian kernel density estimates, for the model parameters obtained through the Bayesian approach (in solid lines) and the ML approach (in dashed lines). From these graphs, we note that the ML estimation leads to more peaked density estimates and therefore underestimates the parameter uncertainty. Furthermore, compared to the Bayesian approach, the ML approach underestimates the values of the components of vector $\boldsymbol{\alpha}$ whereas the components of $\boldsymbol{\beta}$ are overestimated.

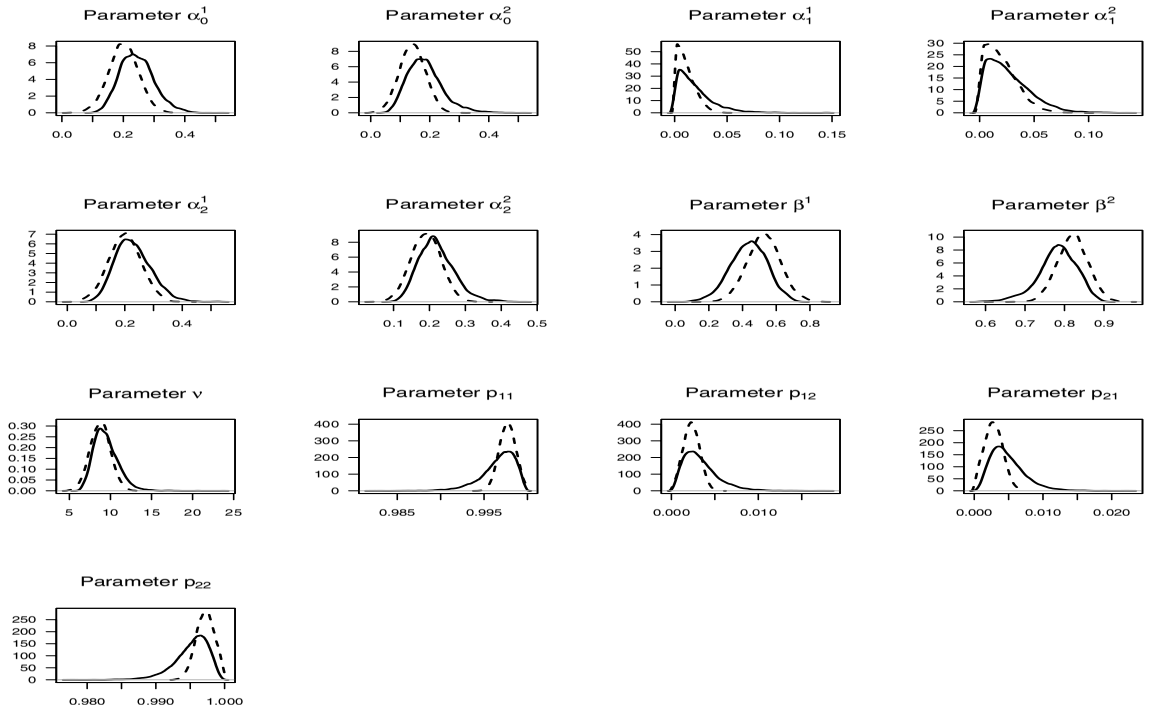


**Fig. 7.9.** Marginal posterior densities of the MS-GJR model parameters and comparison with the asymptotic Normal approximation. Results obtained via the Bayesian approach are given in solid lines while the ML estimates are shown in dashed lines. Gaussian kernel density estimates with bandwidth selected by the "Silverman's rule of thumb" criterion [see Silverman 1986, p.48]. The graphs are based on 10'000 draws from the constrained posterior sample.