Maximum Likelihood Estimators for Multivariate Hidden Markov Mixture Models

Joseph Tadjuidje Kamgaing [*]

University of Kaiserslautern,
Germany

April 15, 2013

Abstract

In this paper we consider a multivariate switching model, with constant states means and covariances. In this model, the switching mechanism between the basic states of the observed time series is controlled by a hidden Markov chain. As illustration, under Gaussian assumption on the innovations and some rather simple conditions, we prove the consistency and asymptotic normality of the maximum likelihood estimates of the model parameters.

1 Introduction

Hidden Markov Models (HMM) are a popular class of models for time series data which, locally within a state, could behave like independent identically distributed (i.i.d.) data, but their statistical properties repeatedly change between states. A prominent example which motivates our approach is portfolio analysis which involves high-dimensional time series data. A simple but wide-spread model for the vector of returns of all assets in a stock portfolio is based on the assumption that the data are independent random Gaussian vectors, e.g., the Delta-Normal model (see Riskmetric [9]), with perhaps constant mean $\boldsymbol{\mu} = 0$, covariance matrix $\boldsymbol{\Sigma}$ and volatility matrix $\boldsymbol{\Sigma}^{1/2}$, respectively. However, if the market environment changes, e.g., if it moves to a more volatile state, the covariance matrix changes too. This behavior can be modeled by a HMM with a finite number, say K, of states represented by the different state means $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$. As such, the HMM model is particularly interesting for analyzing financial returns which are known to exhibit some particularities ("stylized facts", see, e.g., Rydén et al. [10]) such as departure from the normality assumption and existence of dependence between the data. Indeed, it is well known that a hidden Markov mixture model can circumvent both the problem

[*] Email: tadjuidj@mathematik.uni-kl.de

of normality violation as well as that of dependence. Although locally within a state the data could behave as, e.g., i.i.d. Gaussian, the mixture is not necessarily Gaussian. For illustration, if one considers the one dimensional version of the model defined in (2.1), under Gaussian assumption on the residuals, the computation of the skewness or the kurtosis (function of the transition probabilities, states means and variances) indicates that the data from this generating mechanism are not necessarily Gaussian. Moreover, the dependence in time between the data is inherent to the Markovian structure of the hidden switching mechanism. Hence, in this paper we consider a multivariate time series model that can be regarded as a hidden Markov mixture of K different high-dimensional i.i.d. processes that are not necessarily Gaussian.

A prime goal in HMM (Hidden Markov mixture models) is to estimate the model parameters represented by all the $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k = 1, \ldots, K$, as well as the transition probability matrix of the hidden Markov chain. In particular, we focus here on investigating the asymptotic behavior of the maximum likelihood estimators of the model parameters under Gaussian assumption.

2 The Model and the Parameter Estimates

Let $\{Q_t\}$ be a hidden stationary Markov chain with a finite number K of states which controls the data generating mechanism of the observed time series $\{\mathbf{X}_t\}$. Let $A = (a_{ij})_{i,j=1,\ldots,K}$ denote the corresponding transition matrix and $\pi = (\pi_1, \ldots, \pi_K)'$ the stationary distribution of the chain. To simplify notation, we consider a multivariate hidden Markov mixture of processes with constant states trends and covariances, i.e.

$$\mathbf{X}_t = \sum_{k=1}^{K} S_{tk} \left(\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k^{1/2} \boldsymbol{\varepsilon}_t \right) \tag{2.1}$$

where the current state is indicated by

$$S_{tk} = \begin{cases} 1 & \text{if } Q_t = k \\ 0 & \text{otherwise} \end{cases} . \tag{2.2}$$

$\boldsymbol{\Sigma}_k^{-1}$ is the inverse of the covariance matrix $\boldsymbol{\Sigma}_k$ and the innovations $\boldsymbol{\varepsilon}_t$ are assumed to be i.i.d. $\mathcal{N}_d(0, I_d)$ random variables. This serves only as example. Using the method developed here, other distributions of the innovations could be handled. The state variable S_{tk} as defined here indicates that at each time instant one and only one of the hidden states is activated and the remaining $K-1$ are off. The vector of parameters θ combines all the free parameters of the model, i.e. the entries of $\boldsymbol{\mu}_k \in \mathbb{R}^d, \boldsymbol{\Sigma}_k, k = 1, \ldots, K$, as well as $a_{ij} \geq 0, i = 1, \ldots, K, j = 1, \ldots, K-1$. From the Markov chain theory, see e.g. Brémaud [1], A is a stochastic matrix, i.e., $\sum_{j=1}^{K} a_{ij} = 1$, for all $i = 1, \ldots, K$.

In the following, $\Theta \subset \mathbb{R}^p$ denotes the parameter set, and we sometimes write $A = A_\theta$ to stress the dependence of the transition matrix on the parameters.

As stated in the introduction, the model defined in (2.1) belongs to the general framework of hidden Markov mixture of models for which an abundant literature exists, e.g., Cappé et al. [2] or Frühwirth-Schnatter [5] and some of the references therein, just to name a few. Additionally, if we observe that this model also includes states means, it could be regarded as a multivariate extension of the white noises driven by hidden Markov introduced in Francq et al. [4] or a multivariate version of the CHARME model introduced in Stockis et al. [11] with constant states means and covariances.

Below, we shall give conditions on θ which guarantee the existence of a stationary solution to equation (2.1) as well as its geometric ergodicity. Given those conditions are satisfied, we consider the observed process to be sampled from a stationarity and geometrically ergodic mechanism $\{(Q_t, \mathbf{X}_t)\}$, and we assume the starting values $(Q_0, \boldsymbol{X}_0)$ to be generated accordingly to the corresponding stationary distribution. Then, the combined process $\{(Q_t, \mathbf{X}_t)\}_{t=0}^{\infty}$ is a stationary Markov process defined on the product space $\{1, \dots, K\} \times \mathbb{R}^d$.

We always assume that the evolution of the hidden Markov chain does not directly depend on the observed time series, which follows from

A. 2.1. $\{\varepsilon_t\}_{t=0}^{\infty}$ is independent of $\{Q_t\}_{t=0}^{\infty}$.

Then, we have e.g. for $t > 0, k = 1, \dots, K$,

$$\begin{aligned}\mathbb{P}\big(Q_t = k \big| Q_s, \boldsymbol{X}_s, s = 0, \dots, t-1\big) &= \mathbb{P}\left(Q_t = k \big| Q_s, s = 0, \dots, t-1\right) &(2.3)\\ &= \mathbb{P}\left(Q_t = k \big| Q_{t-1}\right) &(2.4)\end{aligned}$$

by the Markov property.

To define the parameter estimates of interest, we first have to introduce some notation. Let $g_\theta(\cdot|k)$ denotes the conditional density of $\mathbf{X}_t$ given $Q_t = k$, which under model (2.1) is the Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{\Sigma}_k$.

Given a sequence $\{y_t\}_{t\in\mathbb{Z}}$ of real vectors and $m, n \in \mathbb{Z},\ m \leq n$, we set $y_m^n = \{y_m, \dots, y_n\}$. For $q_0 \in \{1, \dots, K\}$, we then get the conditional likelihood function as the conditional density of X_1^n given $\boldsymbol{X}_0$ and $Q_0 = q_0$

$$p_\theta(\boldsymbol{X}_1^n | \boldsymbol{X}_0, Q_0 = q_0) = \sum_{q_n=1}^{K} \cdots \sum_{q_1=1}^{K} \prod_{t=1}^{n} a_{q_{t-1},q_t}\ g_\theta(\mathbf{X}_t \mid q_t) \tag{2.5}$$

and the conditional log-likelihood function given $\boldsymbol{X}_0$ and $Q_0 = q_0$

$$\begin{aligned} l_n(\theta, q_0) &= \log p_\theta(X_1^n | \boldsymbol{X}_0, Q_0 = q_0) &(2.6)\\ &= \sum_{t=1}^{n} \log p_\theta(\mathbf{X}_t | X_0^{t-1}, Q_0 = q_0). \end{aligned}$$

with

$$p_\theta(\mathbf{X}_t | X_0^{t-1}, Q_0 = q_0) \quad = \quad \sum_{q_{t-1}=1}^{K} \sum_{q_t=1}^{K} g_\theta(\mathbf{X}_t \mid q_t) a_{q_{t-1},q_t} \mathbb{P}\left(Q_{t-1} = q_{t-1} | X_0^{t-1}, Q_0 = q_0\right).$$

Similarly, the conditional log-likelihood function given only $\boldsymbol{X}_0$

$$l_n(\theta) = \sum_{t=1}^{n} \log \overline{p}_\theta(\mathbf{X}_t | X_0^{t-1}) \tag{2.7}$$

with

$$\overline{p}_\theta(\mathbf{X}_t | X_0^{t-1}) = \sum_{q_{t-1}=1}^{K} \sum_{q_t=1}^{K} g_\theta(\mathbf{X}_t | q_t) a_{q_{t-1},q_t} \mathbb{P}(Q_{t-1} = q_{t-1} | X_0^{t-1}). \tag{2.8}$$

The maximum likelihood estimate

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} l_n(\theta)$$

is the classical likelihood estimate that one will consider, e.g. in the i.i.d. framework without switching mechanism. Indeed, assuming e.g. a Gaussian residuals, one obtains a closed form for the conditional density of $\mathbf{X}_t$. However, assuming a switching mechanism this is not always guaranteed, as one can observe at the begin of Section 5. Therefore, we get interested in

$$\hat{\theta}_{n,q_0} = \arg\max_{\theta \in \Theta} l_n(\theta, q_0),$$

which takes at least into account the switching mechanism. Furthermore, we let $\hat{\theta}_{n,q_0}$ depend on an arbitrary initial value $q_0 \in \{1, \dots, K\}$ which asymptotically will make no difference by Proposition 3.2 below.

3 Asymptotic Properties of the Parameter Estimates

In this section, we state our main results. Under rather weak conditions, we may conclude that the assumptions of Douc et al. [3] are fulfilled for our model and, therefore, make use of their results to derive the asymptotics of the parameter estimates. We assume that the data $\boldsymbol{X}_0, \dots, \boldsymbol{X}_n$ are generated by a multivariate hidden Markov mixture of processes as in (2.1) with unknown parameter θ^*. Moreover, we assume that assumption A.2.1 holds.

A. 3.1. The parameter set Θ is a compact subset of $\mathbb{R}^p$, where p is a function of the number of hidden states and θ^* is an interior point of Θ.

The compactness of the parameter set Θ as well as the assumption on the unknown true parameter value θ^* are quite standard in the literature and will be considered here without any further justification.

A major condition to apply the results of Douc et al. [3] is stationarity, irreducibility and geometric ergodicity of the Markov process $\{(Q_t, \mathbf{X}_t)\}$. We follow the development in Stockis et al. [11], however, with a rather strong assumption on the transition probabilities,

A. 3.2. There exist a_- and a_+ such that

$$0 < a_- \leq a_{ij} \leq a_+ < 1, \text{ for all } i, j = 1, \dots, K; \theta \in \Theta$$

Indeed, the above assumption on the transition probabilities is rather motivated by the proof of the asymptotics of the parameter estimates than by the proof of the geometric ergodicity. In fact, we could have considered weaker assumptions as in [11].
For the covariance matrices let us assume,

A. 3.3. there exist $0 < \delta_l \leq \delta_u < \infty$ such that

$$\delta_l \leq \min_k \lambda_{l,k} \leq \max_k \lambda_{u,k} \leq \delta_u$$

where $\lambda_{l,k}$ is the smallest eigenvalue of the covariance matrix $\mathbf{\Sigma}_k$ and $\lambda_{u,k}$ is its largest eigenvalue.

Assumption 3.3 implies that all the covariance matrices are positive definite and

$$\delta_l^d \le \lambda_{l,k}^d \le |\boldsymbol{\Sigma}_k| \le \lambda_{u,k}^d \le \delta_u^d.$$

Moreover, the existence of δ_u can also be regarded as a consequence of the compactness assumption on the parameter set. Furthermore, $\boldsymbol{\Sigma}_k$ positive definite implies

$$\boldsymbol{\Sigma}_k = \boldsymbol{P}_k^{-1}\boldsymbol{D}_k\boldsymbol{P}_k$$

where $\boldsymbol{D}_k$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{P}_k$ is unitary matrix whose rows comprise an orthonormal basis of eigenvectors of $\boldsymbol{\Sigma}_k$.

From the later observation, it is straightforward to see that the squared Mahalanobis distance, see [8],

$$d^2_{\boldsymbol{\Sigma}_k}(\mathbf{X}_t, \boldsymbol{\mu}_k) = (\mathbf{X}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k)$$

satisfies

$$\frac{1}{\delta_u}\|\mathbf{X}_t - \boldsymbol{\mu}_k\|^2 \le d^2_{\boldsymbol{\Sigma}_k}(\mathbf{X}_t, \boldsymbol{\mu}_k) \le \frac{1}{\delta_l}\|\mathbf{X}_t - \boldsymbol{\mu}_k\|^2. \tag{3.1}$$

Now, using the compactness of Θ, it first follows the existence of $0 \le M < \infty$, such that $\|\mathbf{X}_t - \boldsymbol{\mu}_k\|^2 \le (\|\mathbf{X}_t\| + M)^2$, for all $k = 1, \dots, K$. Therefore, we have

$$\frac{1}{\delta_u}\|\mathbf{X}_t - \boldsymbol{\mu}_k\|^2 \le d^2_{\boldsymbol{\Sigma}_k}(\mathbf{X}_t, \boldsymbol{\mu}_k) \le \frac{1}{\delta_l}(\|\mathbf{X}_t\| + M)^2. \tag{3.2}$$

Given the above assumption we have the following probabilistic result, which induces the stationarity and mixing properties of the observed process. Indeed, the stationarity and mixing properties are key ingredients for deriving the asymptotic properties of the parameter estimates.

Proposition 3.1. *Let $\{\mathbf{X}_t\}$ be generated from model (2.1), and let A.2.1 and A.3.2 to 3.3 hold. It follows*

1. *$\{Q_t\}$ is a strictly stationary, irreducible and aperiodic Markov chain with finite state space $\{1, \dots, K\}$.*
2. *$\{(Q_t, \mathbf{X}_t)\}$ is geometrically ergodic.*
3. *The existence of moments for $\mathbf{X}_t$ follows from the existence of the corresponding moments of $\boldsymbol{\varepsilon}_t$.*

The following result implies that in estimating the model parameter by maximizing the log likelihood, it makes no difference if we assume $Q_0 = q_0$ to be given. The proof is postponed to the technical appendix.

Proposition 3.2. *Consider A.2.1, A.3.1 to A.3.3 hold. It follows*

$$\sup_{\theta} \sup_{1 \le q_0 \le K} \left| \frac{1}{n} l_n(\theta, q_0) - l(\theta) \right| \longrightarrow 0 \qquad a.s. \ as \ n \to \infty$$

For proving consistency, we need a standard identifiability condition for the true parameter vector which is essentially a condition on the parameter set Θ.

A. 3.4. For all $n \geq 1$, θ^* is the unique solution in Θ of

$$\mathbb{E}\left(\log \frac{p_{\theta^*}(X_1^n|\boldsymbol{X}_0)}{p_\theta(X_1^n|\boldsymbol{X}_0)}\right) = 0$$

Giving the above mentioned assumptions and the subsequent results in Proposition 3.1 and 3.2, we are now in position to state and prove the results on the consistency as well as the asymptotic normality of the parameter estimates.

Theorem 3.1. *Let A.2.1 and A.3.1 to A.3.4 hold. Then, for all $q_0 = 1, \ldots, K$,*

$$\lim_{n\to\infty} \hat{\theta}_{n,q_0} = \theta^* \ a.s.$$

where

$$\hat{\theta}_{n,q_0} = \arg\max_{\theta\in\Theta} l_n(\theta, q_0).$$

To formulate the asymptotic normality of the parameter estimates, we have to introduce the notation

$$I(\theta) = -\mathbb{E}_\theta \nabla_\theta^2 \log \overline{p}_\theta(\mathbf{X}_t \mid \boldsymbol{X}_{-\infty}^{t-1}),$$

which does not depend on t for stationary processes. $I(\theta^*)$ is the Fisher information in our model, and we can estimate it consistently as described in the following theorem.

Theorem 3.2. *Consider A.3.1 to A.3.4 and assume, additionally, some moment assumptions, and that $I(\theta^*)$ is positive definite. Then, for all q_0*

$$\frac{1}{n}\nabla_\theta^2 l_n(\hat{\theta}_{n,q_0}, q_0) \longrightarrow I(\theta^*) \ a.s.$$

and

$$\sqrt{n}(\hat{\theta}_{n,q_0} - \theta^*) \longrightarrow \mathcal{N}(0, (I(\theta^*))^{-1}).$$

The later theorem formulates the asymptotic normality of the parameter estimates which is of great importance to the practitioner who wishes e.g., to conduct formal tests of hypothesis and construct confidence interval estimates.

4 Summary and perspectives

In this paper we have proven the consistency and asymptotic normality of the maximum likelihood estimates, for the parameters under Gaussianity assumption on the innovations, for multivariate hidden Markov mixture of AR-ARCH with constant state means and covariances. This should be regarded as an illustration that the theory works and could easily be extended, given different probabilistic type of residuals. However, we always have to check, e.g., that the conditions of Douc et al. [3] hold. Nevertheless, we need to observed that assumption A.3.2 on the transition probabilities could be, perhaps for some applications, rather restrictive. Therefore, it will be worth investigating the asymptotic of the parameter estimates under weaker considerations, for example, allowing some of the transition probabilities to be equal to zero.

5 Technical Appendix

Throughout the whole appendix, we assume that $\{Q_t, \mathbf{X}_t\}$ is a stationary process generated from model (2.1). We first start with some technical lemmas which are needed for proving consistency and asymptotic normality of the maximum likelihood estimates of the parameters. Under model (2.1), the conditional density of $\mathbf{X}_t$ given $\boldsymbol{X}_{t-1}$

$$g_\theta(\mathbf{X}_t \mid \boldsymbol{X}_{t-1}) \quad = \quad \sum_{j=1}^{K} g_\theta(\mathbf{X}_t \mid Q_t = j)\mathbb{P}(Q_t = j \mid \boldsymbol{X}_{t-1})$$

with

$$\begin{aligned}\mathbb{P}(Q_t = j \,|\, \boldsymbol{X}_{t-1}) &= \sum_{i=1}^{K} \mathbb{P}(Q_t = j \,|\, Q_{t-1} = i, \boldsymbol{X}_{t-1})\mathbb{P}(Q_{t-1} = i \mid \boldsymbol{X}_{t-1}) \\ &= \sum_{i=1}^{K} a_{ij}\mathbb{P}(Q_{t-1} = i \,|\, , \boldsymbol{X}_{t-1}).\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{P}(Q_{t-1} = i \,|\, \boldsymbol{X}_{t-1}) &= \mathbb{P}(S_{t-1,i} = 1 \mid \boldsymbol{X}_{t-1}) \\ &= \mathbb{E}(S_{t-1,i} \,|\, \boldsymbol{X}_{t-1})\end{aligned}$$

and using A.2.1 together with the subsequent consequences in equations (2.3) and (2.4), it follows

$$\begin{aligned}\mathbb{P}(Q_{t-1} = i \,|\, \boldsymbol{X}_{t-1}) &= \mathbb{E}(\mathbb{E}(S_{t-1,i} \,|\, S_{t-2}, \boldsymbol{X}_{t-1}) \,|\, \boldsymbol{X}_{t-1}) \\ &= \mathbb{E}(a_{Q_{t-2},i} \,|\, \boldsymbol{X}_{t-1})\end{aligned}$$

Now, putting everything together, we derive

$$g_\theta(\mathbf{X}_t \mid \boldsymbol{X}_{t-1}) \quad = \quad \sum_{i=1}^{K}\sum_{j=1}^{K} g_\theta(\mathbf{X}_t \mid Q_t = j) a_{ij}\mathbb{E}(a_{Q_{t-2},i} \,|\, \boldsymbol{X}_{t-1}) \tag{5.1}$$

where

$$g_\theta(\mathbf{X}_t \mid Q_t = k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k)\right),$$

i.e., the conditional density of the observation given we are in the state k.

Lemma 5.1. *Consider A.2.1and A.3.1 to A.3.3 hold. Then,*

1. *For all $\mathbf{X}_t, \boldsymbol{X}_{t-1} \in \mathbb{R}^d$,*

$$\inf_{\theta\in\Theta} g_\theta(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) > 0, \tag{5.2}$$

and

$$\sup_{\theta\in\Theta} g_\theta(\boldsymbol{X}_t|\boldsymbol{X}_{t-1}) < \infty. \tag{5.3}$$

2. We also obtain,

$$b_+ = \sup_\theta \sup_{\boldsymbol{X}_t, k} g_\theta(\boldsymbol{X}_t \mid k) < \infty \tag{5.4}$$

and

$$\mathbb{E}\left|\log \inf_\theta g_\theta(\boldsymbol{X}_1 \mid \boldsymbol{X}_0)\right| < \infty. \tag{5.5}$$

Proof. 1. Using $\delta_u \leq \min_k \lambda_{l,k}$, for all $\mathbf{X}_t$,

$$g_\theta(\mathbf{X}_t|\boldsymbol{X}_{t-1}) \leq \frac{2K}{(2\pi\delta_u)^{d/2}} < +\infty$$

since $\exp(-\frac{1}{2}u) \leq 1$ for all $u \geq 0$ and $a_{ij} < 1$ for all $i, j = 1, \ldots, K$. On the other hand, by compactness of Θ, we can choose $M > 0$ such that $\|\boldsymbol{\mu}_k\| \leq M, k = 1, \ldots, K$. Then,

$$\begin{aligned} g_\theta(\mathbf{X}_t|\boldsymbol{X}_{t-1}) &\geq \max_{i,j} a_{ij} \frac{a_-}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_j)\right) \\ &\geq \max_j \frac{a_-^2}{(2\pi\delta_u)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\delta_u}\|\mathbf{X}_t - \boldsymbol{\mu}_j\|^2\right) \\ &\geq \frac{a_-^2}{(2\pi\delta_u)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\delta_u}(\|\mathbf{X}_t\| + M)^2\right) > 0 \end{aligned}$$

as $g_\theta(\ |\ )$ sums over positive terms, each of the summand is less than the sum and the first inequality follows since $a_{Q_{t-2},i} \geq a_-$. The second equation follows using $\min_{i,j} a_{ij} \geq a_-$ and (3.1). Finally, the third equation uses the compactness assumption as in (3.2) and the fact that $\exp(-\frac{1}{2}u)$ is decreasing on the positive real line.

2. By definition and moving along the same line of arguments as in the proof of part 1. above, we see that b_+ is trivially dominated by a positive constant. Hence the first part of our assertion holds. For the expectation in the second part, we get from the above development:

$$\frac{2K}{\delta_\sigma} \geq \inf_\theta g_\theta(\boldsymbol{X}_1|\boldsymbol{X}_0) \geq \frac{a_-^2}{(2\pi\delta_u)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\delta_{u^2}}(\|\mathbf{X}_t\| + M)^2\right)$$

Henceforth, there exist a positive constant C such that

$$\left|\log \inf_\theta g_\theta(\boldsymbol{X}_1|\boldsymbol{X}_0)\right| \leq C + \frac{(\|\boldsymbol{X}_1\| + M)^2}{2\delta_u^2}.$$

Using the stationarity and $2nd$ moment assumption for $\mathbf{X}_t$, the assertion follows. □

Proof of Proposition 3.2 and Theorem 3.1:

Under our conditions, the proof of Proposition 3.2 follows along the lines of the proof of Theorem 1 of Stockis et al. [11], which implies that $\{(Q_t, \mathbf{X}_t)\}$ is not only geometrically ergodic, but also irreducible and aperiodic, and every compact set is a petite set. Choosing only the θ for which the drift (one could choose a completely different drift function here) condition is fulfilled, the transition kernel of the combined Markov process $\{(Q_t, \mathbf{X}_t)\}$ is positive Harris recurrent. Therefore, assumption (A2) of Douc et al. [3] is satisfied. Then, Proposition 3.2 follows from going through the proof of Proposition 2 of Douc et al. [3], where we only have to check, if their other assumptions are satisfied too. Our assumptions A.3.2 represents (A1) of Douc et al., (A3) is implied by our Lemma 5.1 below, and (A4) is immediate from the representation (5.1) of $g_\theta(\mathbf{X}_t|\boldsymbol{X}_{t-1})$.

Marking that any stationary process $\{Z_t\}_{t\geq 0}$ can be extended to a two-sided process $\{Z_t\}_{-\infty<t<\infty}$, see e.g. Theorem 4.8 of Krengel [7], our Theorem 3.1 follows from Theorem 1 of Douc et al. [3] once we have checked their conditions. (A1)-(A4) have been discussed already in the previous paragraph. The identifiability condition(A5) follows immediately from our Assumption 3.4. Finally, the required geometric ergodicity follows from Proposition 3.1.

In the next lemma, $\Theta^* \subset \Theta$ denotes an open neighborhood of θ^* contained in Θ which exists by A.3.1. Again, to stress the dependence on the model parameters, we write $A_\theta, a_{kl}(\theta)$ for the transition matrix of $\{Q_t\}$ and its elements.

Lemma 5.2. *Consider A.3.1 to A.3.3 hold. It follows*

(a) for all $k, l \in \{1, \dots, K\}$ and $\boldsymbol{X}_t \in \mathbb{R}^d$, the functions

$$\theta \longmapsto a_{kl}(\theta) \text{ and } \theta \longmapsto g_\theta(\boldsymbol{X}_t|k)$$

are twice continuously differentiable on Θ^*.

(b)

$$\sup_{\theta\in\Theta^*} \sup_{k,l} \|\nabla_\theta \log a_{kl}(\theta)\| < \infty \text{ and } \sup_{\theta\in\Theta^*} \sup_{k,l} \|\nabla^2_\theta \log a_{kl}(\theta)\| < \infty$$

(c)

$$\mathbb{E}\left\{\sup_{\theta\in\Theta^*} \sup_{k} \|\nabla_\theta \log g_\theta(\boldsymbol{X}_1|k)\|\right\} < \infty \text{ and } \mathbb{E}\left\{\sup_{\theta\in\Theta^*} \sup_{k} \|\nabla^2_\theta \log g_\theta(\boldsymbol{X}_1|k)\|\right\} < \infty.$$

Proof. The first part of (a) and (b) follow immediately from the fact that the transition probabilities $a_{kl}(\theta)$ are parameters themselves or, for $l = K$, linear functions of the parameters. Additionally, let us recall that $g_\theta(\mathbf{X}_t|k)$, the conditional density of $\boldsymbol{X}_t$ given that the hidden process is in state k, is a multivariate Gaussian density and in particular

$$\begin{aligned} G_k(\theta) &= \log g_\theta(\mathbf{X}_t|k) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{X}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}d^2_{\boldsymbol{\Sigma}_k}, \end{aligned} \tag{5.6}$$

therefore, the required differentiability of $g_\theta(\mathbf{X}_t|k)$, in (a), follows trivially.

For the rest of the proof of (c), it is enough to investigate and control the first and second order partial derivatives of $G_k(\theta)$ with respect to the entries of states means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$.

It is easy to observe that the partial derivatives which include only the entries of the state means are straightforward and using some matrix algebra could be written in the vector form as (see (7.7) and (3.9) in Harville [6])

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} G_k(\theta) = -2\Sigma_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k), \tag{5.7}$$

and

$$\frac{\partial}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k'} G_k(\theta) = 2\boldsymbol{\Sigma}_k^{-1}. \tag{5.8}$$

However, using the matrix differentiation as exposed in Harville [6], for the more general framework including the covariance matrices component, for all $i, j = 1, \ldots, d$, we have,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \log |\boldsymbol{\Sigma}_k| \quad = \quad tr\left(\boldsymbol{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k\right\}\right). \tag{5.9}$$

Since $\boldsymbol{\Sigma}_k$ is non singular, we derive, using $\boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k = I_d$,

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k^{-1} = -\boldsymbol{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k\right\}\boldsymbol{\Sigma}_k^{-1}. \tag{5.10}$$

Additionally, recalling the symmetric property of $\boldsymbol{\Sigma}_k$ and considering u_j(the jth-column of the identity matrix I_d),

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k = \begin{cases} u_i u_i' & i = j \\ u_i u_j' + u_j u_i' & i > j \text{ or } i < j \end{cases}.$$

Therefore,

$$\frac{\partial^2}{\partial \boldsymbol{\Sigma}_{k,ij} \partial \boldsymbol{\Sigma}_{k,ul}} \boldsymbol{\Sigma}_k = 0_{d \times d}, \text{ for all } i, j, u, l = 1 \ldots, d. \tag{5.11}$$

Using (5.10), we then derive, for all $i, j = 1 \ldots, d$,

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} d^2_{\boldsymbol{\Sigma}_k}(\boldsymbol{X}_t, \boldsymbol{\mu}_k) \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} tr\left((\mathbf{X}_t - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k)\right) \\ &= \frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} tr\left(\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k)(\mathbf{X}_t - \boldsymbol{\mu}_k)'\right) \\ &= tr\left(-\boldsymbol{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k\right\}\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_k)(\mathbf{X}_t - \boldsymbol{\mu}_k)'\right) \end{aligned}$$

Similarly, using (5.9) to (5.11) and allowing for repetitions of the covariances entries, i.e., for all $i, j, u, l = 1 \ldots, d$,

$$\begin{aligned} &\frac{\partial^2}{\partial \boldsymbol{\Sigma}_{k,ij} \partial \boldsymbol{\Sigma}_{k,ul}} \log |\boldsymbol{\Sigma}_k| \\ &= tr\left(\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ul}} \boldsymbol{\Sigma}_k^{-1} \frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k\right) \\ &= tr\left(-\boldsymbol{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ul}} \boldsymbol{\Sigma}_k\right\}\boldsymbol{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \boldsymbol{\Sigma}_{k,ij}} \boldsymbol{\Sigma}_k\right\}\right) \end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^2}{\partial \mathbf{\Sigma}_{k,ij}\partial \mathbf{\Sigma}_{k,ul}} d^2_{\mathbf{\Sigma}}(\boldsymbol{X}_t, \boldsymbol{\mu}_k)\\
&= tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ul}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k)(\mathbf{X}_t-\boldsymbol{\mu}_k)'\right)\\
&+ tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ul}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k)(\mathbf{X}_t-\boldsymbol{\mu}_k)'\right).
\end{aligned}$$

We can now derive the partial derivatives with respect to the covariance entries, i.e., for all $i, j = 1, \ldots, d$,

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}} G_k(\theta) \qquad (5.12)\\
&= -tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\right)\\
&+ tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k)(\mathbf{X}_t-\boldsymbol{\mu}_k)'\right).
\end{aligned}$$

Moreover, the second order partial derivatives with respect to the covariance entries and allowing for repetitions, i.e for all $i, j, u, l = 1, \ldots, d$, are given by

$$\begin{aligned}
&\frac{\partial^2}{\partial \mathbf{\Sigma}_{k,ij}\partial \mathbf{\Sigma}_{k,ul}} G_k(\theta) \qquad (5.13)\\
&= tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ul}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\right)\\
&+tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ul}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k)(\mathbf{X}_t-\boldsymbol{\mu}_k)'\right)\\
&+tr\left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ul}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k)(\mathbf{X}_t-\boldsymbol{\mu}_k)'\right).
\end{aligned}$$

Finally, the cross second order partial derivatives, with respect to the μ_k and the entries of the covariances matrices, are given by

$$\frac{\partial^2}{\partial \boldsymbol{\mu}_k \partial \mathbf{\Sigma}_{k,ij}} G_k(\theta) \quad = \quad 2\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}(\mathbf{X}_t-\boldsymbol{\mu}_k) \qquad (5.14)$$

and, using relation (3.7) in [6],

$$\begin{aligned}
&\frac{\partial^2}{\partial \mathbf{\Sigma}_{k,ij}\partial \boldsymbol{\mu}_k} G_k(\theta) \qquad (5.15)\\
&= -\left[\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\right.\\
&\quad\left. + \left(\mathbf{\Sigma}_k^{-1}\left\{\frac{\partial}{\partial \mathbf{\Sigma}_{k,ij}}\mathbf{\Sigma}_k\right\}\mathbf{\Sigma}_k^{-1}\right)'\right](\mathbf{X}_t-\boldsymbol{\mu}_k).
\end{aligned}$$

We have designed expressions for first and second order partial derivatives. To conclude with the proof, we need to observe that, for the control of the expectations of the norms of some of the above partial derivatives as desired in the statement of the Lemma, e.g. (5.11) where the trace is involved, we make use of

$$(\mathbb{V}\mathrm{ec}\, A)'(\mathbb{V}\mathrm{ec}\, B) = trAB,$$

where $\mathbb{V}\mathrm{ec}$ is the vectorization operator. Furthermore, applying the Cauchy-Schwartz inequality

$$|(\mathbb{V}\mathrm{ec}\, A)'(\mathbb{V}\mathrm{ec}\, B)| \leq \| \mathbb{V}\mathrm{ec}\, A\| \| \mathbb{V}\mathrm{ec}\, B\|.$$

For other partial derivatives, e.g., (5.8), we just need to use the definition of a matrix norm or for some other, e.g., (5.7), remark that for any suitable matrix A and a vector X the induced norm satisfies

$$\|AX\| \leq \|A\|\|X\|.$$

Together with the compactness of the parameter set Θ, we can conclude with the proof of the Lemma by means of some straightforward calculations. □

Before we embark with the proof of asymptotic normality, let us state an additional auxiliary result.

Lemma 5.3. *If A.3.1 to A.3.2 and, additionally,* $\mathbb{E}\|\mathbf{X}_t\|^2 < \infty$ *hold then*

1. *There exists a function* $f_0 : \mathbb{R}^d \longmapsto \mathbb{R}^+$ *satisfying* $\mathbb{E}f_0(\mathbf{X}_t) < \infty$, *such that*
$$\sup_{\theta \in \Theta^*} g_\theta(\boldsymbol{X}_t|k) \leq f_0(\mathbf{X}_t)$$
for all $\boldsymbol{X}_t \in \mathbb{R}^d$.
2. *There exist functions* $f_1, f_2 : \mathbb{R}^d \longmapsto \mathbb{R}^+$ *satisfying* $\mathbb{E}f_i(\mathbf{X}_t) < \infty$, *for* $i = 1, 2$, *such that*
$$\|\nabla_\theta g_\theta(\boldsymbol{X}_t|k)\| \leq f_1(\boldsymbol{X}_t) \quad \text{and } \|\nabla^2_\theta g_\theta(\boldsymbol{X}_t|k)\| \leq f_2(\boldsymbol{X}_t)$$
for all $\boldsymbol{X}_t \in \mathbb{R}^d$.

Proof. We use the notation

$$g_k(\theta) = g_\theta(\mathbf{X}_t \mid k) = \exp(G_k(\theta))$$

1. follows immediately with a constant f_0 from $0 < g_k(\theta) \leq \frac{1}{\delta_\sigma}$ which we have shown above.

For showing 2., let γ_k, ρ_k represent an arbitrary selection of $\boldsymbol{\mu}_k$ or $\boldsymbol{\Sigma}_k$ entries, with repetitions allowed. We have

$$\frac{\partial g_k(\theta)}{\partial \gamma_k} = \frac{\partial G_k(\theta)}{\partial \gamma_k} g_k(\theta), \tag{5.16}$$

and

$$\frac{\partial^2 g_k(\theta)}{\partial \gamma_k \partial \rho_k} = \left(\frac{\partial^2 G_k(\theta)}{\partial \gamma_k \partial \rho_k} + \frac{\partial G_k(\theta)}{\partial \gamma_k} \frac{\partial G_k(\theta)}{\partial \rho_k} \right) g_k(\theta) \tag{5.17}$$

where $G_k(\theta)$ and its partial derivatives are given in the proof of Lemma 5.2. The conclusion follows using similar idea as in the proof of Lemma 5.2, part (c). □

Proof of Theorem 3.2:

Proof. The assertion follows from Theorems 3 and 4 of Douc et al. (2004). We have already discussed in the proof of Theorem 3.1 that their assumptions (A1) to (A5) are fulfilled. The remaining assumptions (A6)-(A8) follow from Lemma 5.2 and 5.3. □

References

[1] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, Newyork, 1999.

[2] Cappé, O., Moulines, E., and Rydén, T. *Inference in hidden Markov models.* Springer Series in Statistics. New York, NY: Springer. , 2005.

[3] Douc, R., Moulines, É., and Rydén, T. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Stat.*, 32(5):2254–2304, 2004.

[4] Francq, C. and Roussignol, M. On white noises driven by hidden markov chains. *Journal of Time Series Analysis*, 18(6):553–578, 1997.

[5] Frühwirth-Schnatter, S. *Finite mixture and Markov switching models.* Springer Series in Statistics. Berlin: Springer., 2006.

[6] Harville, D. A. *Matrix algebra from a statitician's perspective.* Springer , 2008.

[7] Krengel, U. *Ergodic Theorems.* De Gruyter, Berlin, 1985.

[8] Mahalanobis, P. C. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[9] RiskMetrics. Technical document. *JP Morgan. New York, USA.*, 1996.

[10] Rydén, T., Terasvirta, T., and Asbrink, S. Stylized facts of daily return series and the hidden markov model of absolute returns. *Journal of Applied Econometrics*, 13:21–?44, 1998.

[11] Stockis, J.P., Tadjuidje Kamgaing, J., and Franke, J. On geometric ergodicity of CHARME models. *Journal of Time Series Analysis*, 31:141–152, 2010.