NEXIALOG CONSULTING

Working Paper
Markov-Switching Normal-Mixture GARCH

Baye Matar KANDJI and Adrien MISKO

Nexialog Consulting, Paris, France

May 30, 2024

Abstract

We introduce a volatility model in which the conditional volatility is driven by both a Markov switching (MS) sequence and innovations with normal mixture (NM) distributions, called MS-NM-GARCH. The existence of a strictly stationary solution and a second-order stationary solution is discussed. We use the likelihood approach to estimate the parameters of the model and, to our knowledge, establish for the first time the strong consistency of the maximum likelihood estimator (MLE) of a class of MS-GARCH under standard regular conditions. We develop an iterative algorithm based on the Hamilton filter and the Expectation Maximization algorithm to efficiently compute the MLE. Finally, we test our model to real financial data, showcasing its practical relevance.

Keywords: Normal Mixture GARCH, Regime Switching, Hidden Markov Model, Stationarity, Maximum likelihood Estimator, Return and Volatility forecasting

1 Introduction

Since their introduction by Engle (1982) and Bollerslev (1986), the ARCH and GARCH models are among the most widely used to stylize the conditional volatility of financial returns. These models have gained popularity due to their ability to reproduce important empirical characteristics observed in real financial returns, such as volatility clustering and memory. They are also appreciated for their parsimonious structure, in which the conditional variance is a linear combination of the squares of past volatilities and returns. Several variants of the ARCH and GARCH models have been developed to address certain gaps in these models. These include TGARCH of Zakoian (1994), introduced to capture the asymmetric effects observed on the conditional distribution depending on the sign of the past returns; The ARCH(∞) of Robinson (1991) is designed to capture the long-memory effect. The GARCH-X models allow to have time-variying coefficients depending on observed covariates, such as macroeconomic variables. The class of Markov-Switching GARCH (MS-GARCH), introduced by Hamilton and Susmel (1994) and Cai (1994) enables to incorporate regime changes in the volatility process induced by phases of economic/financial crises, recessions, or depressions.

Most of these models consider a single component of conditional volatility. However, several applied researches, see Lee and Engle (1993), Engle and Rangel (2008), Engle et al. (2013), suggest that multi-component volatility models are more effective to capture complex dynamics such as long memory effects by enabling improved long-term volatility forecasts. Moreover, classical GARCH considers sequence of independent and identically distributed (iid) innovations, where the marginal distribution is in practice assumed to be Gaussian. This condition implies some restrictive features like null conditional skewness and excess kurtosis whereas it is well known that the conditional distribution of daily financial returns is Leptokurtic, see for instance Nelson (1996) and Johnston and Scott (2000). To address this shortcoming and leverage the advantages of multi-components modeling, Haas et al. (2004a) and Alexander and Lazar (2006) independently

introduce the normal mixture-GARCH model (NM-GARCH). In this model, the conditional distribution is a mixture of Normal laws, incorporating multiple component volatilities. Their authors conclude that this class of models can successfully reproduce the excess kurtosis and is particularly suited to model and predict the conditional density and volatility of financial returns, such as exchange rates. However, this model is not structurally designed to take account of regime changes. The MS-GARCH constitutes another class of GARCH models in which the conditional distribution is a mixture distribution with time-varying mixture coefficients, allowing regime changes. However, the inference of these models is very difficult due to the path dependence of the likelihood on the latent factor. In practice, only MS-ARCH models are considered to avoid this dependence, leading to a persistence loss. To overcome this dependency and gain persistence, Haas et al. (2004b) introduce a version of the MS-GARCH model without path dependence in the conditional volatilities. However, in the case of a known presence of different regimes with low probabilities of state change, such as in stock markets, the MS-GARCH could potentially have difficulty to arbitrate between capturing the no-normality of the innovations and effectively filtering states by conditionally considering a classical GARCH over a long period. Moreover, Haas et al. (2004b) note that the structure of the NM-GARCH, unlike the MS-GARCH, allows to consider different components means, that create skewness, without altering the first-order dynamics of the process.

We introduce in this paper, a Markov-switching Normal-mixture GARCH model (MS-NM-GARCH) in which the conditional volatility is driven by both a Markov switching sequence and innovations with normal-mixture distributions. We use the maximum likelihood approach to estimate our parameters, classically used for the inference of GARCH models. The theoretical properties of the maximum likelihood estimator (MLE) of a non-regime switching GARCH are well known, see Francq and Zakoian (2004) and Berkes et al. (2003) for the classical GARCH, Lee and Lee (2009) establish the asymptotic properties of the MLE of the NM-GARCH. For the class of pure MS-GARCH, the MLE theoretical properties have never been established to our knowledge so far. The difficulty in establishing these properties lies in the path dependence on the latent factor or in the non-known effect of the initial value in the quasi-likelihood. For MS-ARCH which do not present these problems, the consistency of the MLE is established by Francq et al. (2001) and asymptotic result for a more general model included the MS-ARCH can be found in Douc et al. (2004). In this paper, we establish the strong consistency of the MS-NM-GARCH MLE, this result is a first of this type for the class of MS-GARCH models.

The paper is organized as follows. In Section 2, we introduce the model and discuss the existence of a stationary solution. Section 3 is dedicated to its calibration with MLE and Hamilton filter. Finally, section 4 deals with empirical results. We decided to put all the proofs in appendix A to make the read easier.

2 The Model

2.1 Description

We say that a random variable follows a normal mixture law if its distribution admits a density of the form

$$\Phi(x) = \sum_{i=1}^{q} m_i \varphi_i(x) \quad \sum_{i=1}^{q} m_i = 1 \quad \varphi_i(x) = \varphi\left(x; \mu_i, \sigma_i^2\right)$$

where $[m_1, m_2, \ldots, m_q]$ is the positive mixing law, and φ denotes the normal density function. We will denote this as $X \sim \text{NM}\left(m_1, \ldots, m_q; \mu_1, \ldots, \mu_q; \sigma_1^2, \ldots, \sigma_q^2\right)$.

Remark 2.1. *Let η and δ be two independent random variables with respective laws: the standard Gaussian distribution, and the distribution with support $\{1, \ldots, q\}$ with respective probabilities $\{m_1, \ldots, m_q\}$. One can remark that $\mu_\delta + \sigma_\delta \eta$ follows a $\text{NM}\left(m_1, \ldots, m_q; \mu_1, \ldots, \mu_q; \sigma_1^2, \ldots, \sigma_q^2\right)$ distribution. Conversely, by Lemma A.1, every random variable X with a normal mixture distribution can be decomposed in the latter form with $\delta(X) := \delta$ and $\eta(X) := \eta$.*

We define the d-q Markov Switching Normal Mixture GARCH process MS(d)-NM(q)-GARCH as follows:

$$\begin{cases} r_t \mid I_{t-1} \sim \mathrm{NM}\left(m_{1,\tilde{\Delta}_t}, \ldots, m_{q,\tilde{\Delta}_t}; \mu_1, \ldots, \mu_q; \sigma^2_{1,t}, \ldots, \sigma^2_{q,t}\right), \ \sum_{i=1}^q m_{i,\tilde{\Delta}_t} = 1 \\ \sigma^2_{i,t} = \omega_i + \alpha_i (r_{t-1} - \mu_{t-1})^2 + \sum_{j=1}^q \beta_{i,j} \sigma^2_{j,t-1}, \ \text{for } i = 1, 2, \ldots, q. \end{cases} \tag{2.1}$$

where I_t represents the information set available at time t, $\mu_t = \mu_{\delta(r_t | I_{t-1})}$, $(\tilde{\Delta}_t)$ is a irreducible and aperiodic I_{t-1} measurable Markov chain with states $\{1, 2, \ldots, d\}$. Let $\tilde{\boldsymbol{P}} = (p_{i,j})_{i,j} = \mathbb{P}(\tilde{\Delta}_t = j \mid \tilde{\Delta}_t = i)$ be the transition matrix and $(m_{i,j})_{i,j} = \mathbb{P}(\delta_t = i \mid \tilde{\Delta}_t = j)$ be the mixture matrix.

$\boldsymbol{\omega} = (\omega_i)_i, \ \boldsymbol{\alpha} = (\alpha_i)_i, \ (\beta_{i,j})_{i,j}, (p_{i,j})_{i,j}, \boldsymbol{m}(j) = (m_{i,j})_i$ are positive and $\omega_1 < \omega_2 < \cdots < \omega_q$

This more general model adds a time varying mixture depending on a Hidden Markov process, contrary to the normal mixture GARCH(1,1) considered by Haas et al. (2004a) and Alexander and Lazar (2006). Model (2.1) includes several conditional volatility models. The Normal Markov Switching ARCH is obtained when $q = d$, $m_{i,i} = 1$, $\mu_i = 0$ and $\beta_{i,j} = 0$ for all i, j. If $d = 1$, we obtain the NM-GARCH of Alexander and Lazar (2006), and if $q = d$, $m_{i,i} = 1$, and $\mu_i = 0$ for all i, the model is reduced to the MS-GARCH of Haas et al. (2004b).

Remark 2.2. *The main idea of this new model is to use the NM feature to better fit the structure of the data (allowing excess of Kurtosis and fat tails) and use the MS feature to monitor the structure change with time (especially in high volatility period when the Kurtosis may increase drastically during a short period) and amplify the NM's effect. As we'll see in the next sections, state 1 should reflect strong long memory effect observed in steady period and state q low memory and fast adaptation to recent data as observed in crisis period; the other states act as intermediate levels between these two. The transition matrix reflects the persistence of each state, in other words, how long the model has to stay in a high kurtosis period without new information. The matrix M monitors the level of kurtosis in each state.*

2.2 Stationarity

In this section, we derive sufficient and necessary conditions to guarantee strict and second order stationarity.

To study the existence of a strictly stationary solution of model (2.1), we reformulate it into a more suitable equivalent form. For all t, let $\delta_t = \delta(r_t \mid I_{t-1})$ and $\eta_t = \eta(r_t \mid I_{t-1})$. By Remark 2.1, it is not difficult to show that conditionally to $\tilde{\Delta}_t$, δ_t is an independent process, its conditional distribution has support $\{1, \ldots, q\}$ and η_t is iid with standard Gaussian marginal distribution, independent of $\tilde{\Delta}_t$ and δ_t. It follows that the process ϵ_t defined by: $\epsilon_t = r_t - \mu_{\delta_t}$ for all t, verify the following model:

$$\begin{cases} \epsilon_t = \sigma_{\delta_t, t} \eta_t \\ \sigma^2_{i,t} = \omega_i + \alpha_i \epsilon^2_{t-1} + \sum_{j=1}^q \beta_{i,j} \sigma^2_{j,t-1}, \ \text{for } i = 1, 2, \ldots, q. \end{cases} \tag{2.2}$$

Thus, the existence of stationary of (2.1) and (2.2) are equivalent and we now focus on the process ϵ_t.

2.2.1 Strict Stationarity

Defining $(b_t)_{i,j} = \alpha_i \eta^2_{t-1} \mathbb{1}_{\delta_{t-1}=j} + \beta_{i,j}$ for all $i, j \leq q$, we have $\sigma^2_{i,t} = \omega_i + \sum_{j=1}^q (b_t)_{i,j} \sigma^2_{j,t-1}$ for all i. Thus we can rewrite (2.2) in the vector form as

$$\boldsymbol{\sigma}^2_t = \boldsymbol{\omega} + \mathbf{B}_t \boldsymbol{\sigma}^2_{t-1}$$

where

$$\boldsymbol{\sigma}^2_t = \left(\sigma^2_{t,1}, \ldots, \sigma^2_{t,q}\right)', \quad \boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)', \quad \mathbf{B}_t = \left((b_t)_{i,j}\right)_{i,j}$$

Since $\mathbb{E} \log^+ \|\mathbf{B}_0\|$ is finite, we can define

$$\gamma := \inf_{n\geq 1} \frac{1}{n}\mathbb{E}\log\|\mathbf{B}_n\mathbf{B}_{n-1}\cdots\mathbf{B}_1\|.$$

Since (δ_t, η_t) is stationary and ergodic, the definition of γ is independent of the chosen norm due to the equivalence of norms and the Kingman subadditive ergodic theorem, see Kingman (1973), which states that:

$$\boldsymbol{\gamma} = \lim_{n\to+\infty} \frac{1}{n}\mathbb{E}\log\|\mathbf{B}_n\mathbf{B}_{n-1}\cdots\mathbf{B}_1\| = \lim_{n\to+\infty} \frac{1}{n}\log\|\mathbf{B}_n\mathbf{B}_{n-1}\cdots\mathbf{B}_1\| \quad a.s. \tag{2.3}$$

For all $t \geq 0$, let $\mathbf{B}_t^{(0)}$ the identity matrix and $\mathbf{B}_t^{(n)} = \mathbf{B}_t \circ \cdots \circ \mathbf{B}_{t-n+1}$ for all $n \geq 1$. Consider the following condition

$$\mathbb{P}(\inf_i\{(\sum_{k=0}^{+\infty}\mathbf{B}_0^{(k)}\boldsymbol{\omega})_i,\ i \leq q\} = 0) < 1. \tag{2.4}$$

Theorem 2.1. *If $\gamma < 0$ under (2.4), model (2.2) (and thus model (2.1)) admits a (unique) strictly stationary (and ergodic) solution. The vector component volatilities is given by*

$$\boldsymbol{\sigma}_t^2 = \sum_{k=t}^{+\infty} \boldsymbol{B}_t^{(k)}\boldsymbol{\omega}, \quad \textit{for all } t.$$

Conversely, if model (2.2) (or model (2.1)) admits a positive strictly stationary solution under (2.4), then $\gamma < 0$.

A process is strictly stationary if every finite subsequence has the same distribution as any of its shifts. This property is often necessary for the statistical inference of econometric models, such as GARCH, as in our setup.

Proof. Direct consequence of Kandji (2024, Theorem 3.1) □

It is well known that γ seems to be impossible to compute explicitly, but estimating it through computer simulations using (2.3) is feasible but expensive. For all squared matrice $\boldsymbol{M}$, we denote $\rho(\boldsymbol{M})$ the spectral radius of $\boldsymbol{M}$. Let $\boldsymbol{\beta} = (\beta_{i,j})_{i,j}$. Since $0 \leq \boldsymbol{\beta} \leq \mathbf{B}_t$ for all t, a direct consequence of the condition $\gamma < 0$ is that $\rho(\boldsymbol{\beta}) < 1$. In practice, it's not necessary to constrain the model's calibration in the space of stationarity parameters but only on $\boldsymbol{\beta}$, along with other regularity conditions not linked to stationarity, to sufficiently guarantee the consistency of the likelihood estimator.

For the classical MS-GARCH model, Francq et al. (2001) established the result of Theorem 2.1 under the assumption that $\inf_i \omega_i > 0$. It is easy to see that this condition implies (2.4), but the converse is not necessarily true; an easily checked example is that: if $\boldsymbol{\beta}$ is irreducible and $\sup_i \omega_i > 0$, even if $\inf_i \omega_i = 0$, then (2.4) hold. Furthermore, note that Theorem 2.1 pertains strictly stationary solutions without additional constraint, in contrast to Alexander and Lazar (2006) and Haas et al. (2004b), which only provide conditions for second-order stationary solutions for the NM-GARCH and MS-GARCH.

2.2.2 Second Order Stationarity

Let us denote $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \cdots, \tilde{\pi}_d)$ the stationary distribution of $\tilde{\Delta}_t$. Let us define $\tilde{b}_t(\tilde{\Delta}_t)_{i,j} = \alpha_i m_{j,\tilde{\Delta}_t} + \beta_{i,j}$, $\tilde{\boldsymbol{B}}(\tilde{\Delta}_t) = \left(\tilde{b}_t(\tilde{\Delta}_t)_{i,j}\right)_{i,j}$ and let the $dq \times dq$ and $d \times dq$ matrices:

$$\boldsymbol{Q} = \begin{pmatrix} p_{1,1}\tilde{\boldsymbol{B}}(1) & p_{2,1}\tilde{\boldsymbol{B}}(1) & \cdots & p_{d,1}\tilde{\boldsymbol{B}}(1) \\ p_{1,2}\tilde{\boldsymbol{B}}(2) & p_{2,2}\tilde{\boldsymbol{B}}(2) & \cdots & p_{d,2}\tilde{\boldsymbol{B}}(2) \\ \vdots & & & \\ p_{1,d}\tilde{\boldsymbol{B}}(d) & p_{2,d}\tilde{\boldsymbol{B}}(d) & \cdots & p_{d,d}\tilde{\boldsymbol{B}}(d) \end{pmatrix}, \text{ and } \boldsymbol{\psi} = \begin{pmatrix} p_{1,1}\boldsymbol{m}(1) & p_{2,1}\boldsymbol{m}(1) & \cdots & p_{d,1}\boldsymbol{m}(1) \\ p_{1,2}\boldsymbol{m}(2) & p_{2,2}\boldsymbol{m}(2) & \cdots & p_{d,2}\boldsymbol{m}(2) \\ \vdots & & & \\ p_{1,d}\boldsymbol{m}(d) & p_{2,d}\boldsymbol{m}(d) & \cdots & p_{d,d}\boldsymbol{m}(d) \end{pmatrix}.$$

Theorem 2.2. *If $\rho(\boldsymbol{Q}) < 1$ then model (2.1) admits a (unique) second order strictly stationary (and ergodic) solution. The variance of r_t is then given by*

$$\operatorname{var}(r_t) = \mathbf{1}'_{(d)}\boldsymbol{\psi}\left(I_{dq} - \boldsymbol{Q}\right)^{-1}\boldsymbol{z} + \sum_{k=1}^{d}\mu(k)^2\tilde{\pi}_k - \left(\sum_{k=1}^{d}\mu(k)\tilde{\pi}_k\right)^2 \tag{2.5}$$

where $\mathbf{1}'_{(n)} = (1,\dots,1)$ and $\boldsymbol{z} = (\tilde{\pi}_1\boldsymbol{\omega}',\dots,\tilde{\pi}_d\boldsymbol{\omega}')'$ are respectively $1\times n$ and $dq\times 1$ matrices, I_n the $n\times n$, identity matrix, $\mu(k) = \sum_{i=1}^{q} m_{i,k}\mu_i$ and $\boldsymbol{m}(k) = (m_{i,k})_i$.

Conversely, suppose that $\boldsymbol{Q}$ is irreducible and that $\sup_i \omega_i > 0$; if model (2.1) admits a positive second order strictly stationary solution then $\rho(\boldsymbol{Q}) < 1$.

A second-order stationary sequence is a strictly stationary sequence with a finite second-order moment. The existence of a second-order moment can be used to measure the overall variability of our series as well as to evaluate the autocorrelations.

Proof. Stated in A.2 □

Remark 2.3. *The result of Theorem 2.2 aligns with the properties established by Francq et al. (2001) for the classical MS-GARCH without component means. It is also easy to see that if $q = d$, $\mu_i = 0$, and $m_{i,i} = 1$ for all $i \le q$, then $\mathbf{1}'_{(q)}\boldsymbol{\psi} = vec(\mathbf{P})'$, and the last two terms of Eq. (2.5) are zero, we recover the variance formula of Haas et al. (2004b, Eq. 13) for their MS-GARCH.*

Unlike the condition imposed for strict stationarity, second-order stationarity is, in practice, very easy to check using the condition $\rho(\boldsymbol{Q}) < 1$, given in practice d and q are small. Note also that in parameter estimation, this constraint could be added to obtain a second-order stationary model without significant additional cost.

3 Calibration of Parameters

3.1 State space redefinition

We reformulate the model in another equivalent form in which the estimator is more practical to use. Let $\Delta_t = (\tilde{\Delta}_t, \delta_t) = (\Delta_{t,1}, \Delta_{t,2})$. The idea is to reduce the model to one hidden factor with dq states.

From (2.2) we deduct that :

$$\begin{cases} r_t = \mu_{\Delta_t} + \sigma_{\Delta_t}\eta_t, \quad \mu_{\Delta_t} = \mu_{\Delta_{t,2}}\ , \quad \sigma_{\Delta_t} = \sigma_{\Delta_{t,2},t} \\ \sigma^2_{i,t} = \omega_i + \alpha_i(r_{t-1} - \mu_{\Delta_{t-1}})^2 + \sum_{j=1}^{q}\beta_{i,j}\sigma^2_{j,t-1}, \text{ for } i = 1,\dots,q \end{cases} \tag{3.1}$$

Here r_t depends on $\Delta_{t,1}$ through $\Delta_{t,2}$ and Δ_t is a Markov chain because:

$$\begin{aligned} &\mathbb{P}\left\{(\Delta_{t,1},\Delta_{t,2}) = (i_t,j_t) \mid (\Delta_{k,1},\Delta_{k,2}) = (i_k,j_k) :\ k < t\right\} \\ &= \mathbb{P}\left\{\Delta_{t,2} = j_t \mid \Delta_{t,1} = i_t, (\Delta_{k,1},\Delta_{k,2}) = (i_k,j_k) :\ k < t\right\} \\ &\qquad \times \mathbb{P}\left\{\Delta_{t,1} = i_t \mid (\Delta_{k,1},\Delta_{k,2}) = (i_k,j_k) :\ k < t\right\} \quad [1] \\ &= \mathbb{P}\left\{\Delta_{t,2} = j_t \mid \Delta_{t,1} = i_t\right\} \times \mathbb{P}\left\{\Delta_{t,1} = i_t \mid \Delta_{t-1,1} = i_{t-1}\right\} \\ &= \mathbb{P}\left\{\Delta_{t,1} = i_t,\ \Delta_{t,2} = j_t \mid \Delta_{t-1,1} = i_{t-1}\right\} \\ &= \mathbb{P}\left\{(\Delta_{t,1},\Delta_{t,2}) = (i_t,j_t) \mid (\Delta_{t,1},\Delta_{t,2}) = (i_{t-1},j_{t-1})\right\}. \end{aligned}$$

Using the second-to-last equality of the above relation, we deduced that the transition probabilities of Δ_t is given by the entries $\mathbb{P}\left\{\Delta_t = (i,j) \mid \Delta_{t-1} = (k,l)\right\} = m_{j,i}p_{k,i}$. Besides, we can note that Δ_t is a Markov process with dq states. Indeed, by considering a bijection $\kappa \mid \kappa(i,j) = (i-1)q+j$, we can change the state space from $\{(i,j) :\ 1 \le i,j \le d,q\}$ to $\{i :\ 1 \le i \le dq\}$.

[1] Given $\tilde{\Delta}_t$ and $\tilde{\Delta}_{t-1}$ respectively, the variables $(\tilde{\Delta}_k,\delta_k)$ and $(\tilde{\Delta}_{k-1},\delta_k)$ for $k < t$ do not convey information on $\delta_t = j_t$ and $\tilde{\Delta}_t$ respectively.

The transition matrix $\boldsymbol{P}$ and the (unique) stationary distribution π' of Δ_t are respectively :

$$\mathrm{P} = \begin{pmatrix} p_{1,1}\boldsymbol{C}(1) & p_{1,2}\boldsymbol{C}(2) & \cdots & p_{1,d}\boldsymbol{C}(d) \\ p_{2,1}\boldsymbol{C}(1) & p_{2,2}\boldsymbol{C}(2) & \cdots & p_{2,d}\boldsymbol{C}(d) \\ \vdots & & & \\ p_{d,1}\boldsymbol{C}(1) & p_{d,2}\boldsymbol{C}(2) & \cdots & p_{d,d}\boldsymbol{C}(d) \end{pmatrix}, \quad \pi' = \begin{pmatrix} \tilde{\pi}_1 \boldsymbol{m}(1)' \\ \tilde{\pi}_2 \boldsymbol{m}(2)' \\ \vdots \\ \tilde{\pi}_d \boldsymbol{m}(d)' \end{pmatrix} \tag{3.2}$$

Where for all $k \geq q$, $\boldsymbol{C}(k)$ is the q^2 matrix where all its rows are identical to $\boldsymbol{m}(k) = (m_{i,k})_i$

3.2 MLE derivation and convergence

In real financial data, the component means are estimated to be closer to zero, see (Alexander and Lazar, 2006, table 2). Additionally, most conditional volatility models consider them null. We chose a less restrictive approach in our model and assume them constant, denoted as μ. Thus, we limit the number of parameters without losing too much information and avoid the problem of likelihood path dependence of the latent factor. The parameters of the model (2.1) are then:

$$\boldsymbol{\theta} = (\mu, (\omega_i, \alpha_i)_i, (\beta_{i,j})_{i,j}, (p_{i,j})_{i,j}, (m_{i,j})_{i,j}) \,.$$

Let τ be the projection in the second axis of the reciprocal function of κ, reflecting the values of $\Delta_{t,2}$, and let $(r_1, \ldots, r_n)$ be a realization of (2.1). Given the initial values $\sigma_{1,1} = \sigma_{2,1} = \ldots = \sigma_{q,1} = 0$ and an initial distribution $\boldsymbol{\pi}_0 > 0$, the conditional likelihood $L_\theta(r_1, \ldots, r_n)$ is given by summing over all the possible $(e_1, \ldots, e_n)$ of the Markov chain, where the e_i belong to $\mathcal{E} = \{1, \ldots, dq\}$:

$$L_\theta(r_1, \ldots, r_n) = \sum_{(e_1, \ldots, e_n) \in \mathcal{E}^n} \boldsymbol{\pi}_0(e_1) \left\{ \prod_{t=2}^{n} P_{e_{t-1}, e_t} \right\} \left\{ \prod_{t=1}^{n} \phi_{e_t}(r_1, \ldots, r_t) \right\} \tag{3.3}$$

where $\phi_{e_t}(r_1, \ldots, r_t) = \frac{1}{(2\pi)^{1/2}\sigma_{\tau(e_t),t}} \exp\left\{ -\frac{(r_t - \mu)^2}{2\sigma^2_{\tau(e_t),t}} \right\}$ and the $(\sigma_{i,t})_i$ are defined recursively by

$$\boldsymbol{\sigma}_t^2 = \boldsymbol{c}_t + \boldsymbol{B}\boldsymbol{\sigma}_{t-1}^2, \tag{3.4}$$

where $\boldsymbol{\sigma}_1 = (\sigma_{1,1}, \sigma_{2,1}, \ldots, \sigma_{q,1})'$, and for all $t > 1$,

$$\boldsymbol{\sigma}_t^2 = \begin{pmatrix} \sigma_{1,t}^2 \\ \sigma_{2,t}^2 \\ \vdots \\ \sigma_{p,t}^2 \end{pmatrix}, \quad \boldsymbol{c}_t = \begin{pmatrix} \omega_1 + \alpha_1(r_{t-1} - \mu)^2 \\ \omega_2 + \alpha_2(r_{t-1} - \mu)^2 \\ \vdots \\ \omega_p + \alpha_p(r_{t-1} - \mu)^2 \end{pmatrix}, \quad B = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,q} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,q} \\ \vdots & & & \\ \beta_{q,1} & \beta_{q,2} & \cdots & \beta_{q,q} \end{pmatrix}.$$

Denoted by Θ the space of parameter, the maximum likelihood estimator is defined as any measurable solution of

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} L_{\boldsymbol{\theta}}(r_1, \ldots, r_n)$$

We show in A.3 that under regular assumptions, $\widehat{\boldsymbol{\theta}}_n$ converges almost surely to $\boldsymbol{\theta}_0$ as $n \to \infty$. Unfortunately, the formula 3.3 cannot be used for calibration due to the considerable number of terms in the sum. Various techniques have been developed in existing literature to calculate the likelihood. For simplicity, we chose one of the most popular: the Hamilton Filter coupled with the Expectation Maximisation technique.

3.3 Computation by the Hamilton Filter

In the subsequent sections, we adapt the procedure described in (Francq and Zakoian, 2019, Section 12.2.1). Let us denote $b := dq$, and let

$$\boldsymbol{\pi}_{t|t} = \begin{pmatrix} \mathbb{P}(\Delta_t = 1 \mid r_t, \ldots, r_1) \\ \vdots \\ \mathbb{P}(\Delta_t = b \mid r_t, \ldots, r_1) \end{pmatrix}, \quad \boldsymbol{\pi}_{t|t-1} = \begin{pmatrix} \mathbb{P}(\Delta_t = 1 \mid r_{t-1}, \ldots, r_1) \\ \vdots \\ \mathbb{P}(\Delta_t = b \mid r_{t-1}, \ldots, r_1) \end{pmatrix}$$

By Bayes' theorem, we have

$$\pi_{t|t}(i) = \mathbb{P}(\Delta_t = i \mid r_t, \ldots, r_1) = \frac{f_t(r_t \mid \Delta_t = i, r_{t-1}, \ldots, r_1)\, \mathbb{P}(\Delta_t = i \mid r_{t-1}, \ldots, r_1)}{f_t(r_t \mid r_{t-1}, \ldots, r_1)}$$

where $f_t(r_t \mid r_{t-1}, \ldots, r_1) = \sum_{i=1}^{b} f_t(r_t \mid \Delta_t = i, r_{t-1}, \ldots, r_1)\, \mathbb{P}(\Delta_t = i \mid r_{t-1}, \ldots, r_1)$

$$f_t(r_t \mid \Delta_t = i, r_{t-1}, \ldots, r_1) = \frac{1}{(2\pi)^{1/2}\sigma_{\tau(i),t}} \exp\left\{-\frac{(r_t - \mu)^2}{2\sigma^2_{\tau(i),t}}\right\}. \tag{3.5}$$

Let us denote by $\odot$ the element-by-element Hadamard product of matrices and let

$$\boldsymbol{f_{t|t-1}} = (f_t(r_t \mid \Delta_t = i, r_{t-1}, \ldots, r_1))'_{i \leq b} \tag{3.6}$$

Starting from an initial value $\boldsymbol{\pi}_{1|0} = \boldsymbol{\pi}$ (the stationary law), we can recursively compute

$$\begin{cases} \boldsymbol{\pi}_{t|t} = \dfrac{\boldsymbol{\pi}_{t|t-1} \odot \boldsymbol{f_{t|t-1}}}{\mathbf{1}' \{\boldsymbol{\pi}_{t|t-1} \odot \boldsymbol{f_{t|t-1}}\}} \\ \boldsymbol{\pi}_{t+1|t} = \mathrm{P}' \boldsymbol{\pi}_{t|t} \end{cases} \tag{3.7}$$

for $t = 1, \ldots, n$

We deduce the conditional log-likelihood

$$\log L_\theta = \sum_{t=1}^{n} \log f_t(r_t \mid r_{t-1}, \ldots, r_1) \tag{3.8}$$

where

$$f_t(r_t \mid r_{t-1}, \ldots, r_1) = \mathbf{1}' \{\boldsymbol{\pi}_{t|t-1} \odot \boldsymbol{f_{t|t-1}}\}. \tag{3.9}$$

3.4 Expectation-Maximisation

Here we consider that the initial distribution $\boldsymbol{\pi}$ is not necessarily the stationary distribution. In the **EM** algorithm, $\boldsymbol{\pi}$ is an additional parameter to be estimated.

Step E: Suppose that an estimator $\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\pi}^{(k)}\right)$ of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ is available. It seems sensible to approximate the unknown log-likelihood by its expectation given the observations $(r_1, \ldots, r_n)$, evaluated under the law parameterised by $\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\pi}^{(k)}\right)$. Let $\hat{\boldsymbol{\theta}} = (\mu, (\omega_i, \alpha_i)_i, (\beta_{i,j})_{i,j})$ be the parameters involved in the GARCH part and $(m_{i,j})_{i \leq q, j \leq d}$ the mixture parameters. We get the criterion

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\pi}^{(k)}\right) = E_{\boldsymbol{\theta}^{(k)}, \boldsymbol{\pi}^{(k)}} \{\log L_{\boldsymbol{\theta}, \boldsymbol{\pi}}(r_1, \ldots, r_n, \Delta_1, \ldots, \Delta_n) \mid r_1, \ldots, r_n\}$$

It is shown that maximizing Q improves the likelihood at each iteration and, under regularity conditions, the algorithm converges to a local extremum, dependent on the initial parameter, see Dempster et al. (1977).

By conditioning, one has

$$\begin{aligned} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1,\ldots,r_n,\Delta_1,\ldots,\Delta_n\right) &= L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1,\ldots,r_n \mid \Delta_1,\ldots,\Delta_n\right) L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\Delta_1,\ldots,\Delta_n\right) \\ &= L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1,\ldots,r_n \mid \Delta_1,\ldots,\Delta_n\right) \\ &\times L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\delta_1,\ldots,\delta_n \mid \tilde{\Delta}_1,\ldots,\tilde{\Delta}_n\right) L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\tilde{\Delta}_2,\ldots,\tilde{\Delta}_n \mid \tilde{\Delta}_1\right) L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\tilde{\Delta}_1\right) \end{aligned}$$

and

$$\begin{aligned} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1,\ldots,r_n \mid \Delta_1,\ldots,\Delta_n\right) &= L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1,\ldots,r_n \mid \delta_1,\ldots,\delta_n\right) \\ &= L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_1 \mid \delta_1\right) \prod_{t=2}^{n} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_t \mid \delta_t, r_1,\ldots,r_{t-1}\right) \\ L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\delta_1,\ldots,\delta_n \mid \tilde{\Delta}_1,\ldots,\tilde{\Delta}_n\right) &= \prod_{t=1}^{n} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\delta_t \mid \tilde{\Delta}_t\right) \\ L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\tilde{\Delta}_2,\ldots,\tilde{\Delta}_n \mid \tilde{\Delta}_1\right) &= \prod_{t=2}^{n} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\tilde{\Delta}_t \mid \tilde{\Delta}_{t-1}\right). \end{aligned}$$

Since

$$\begin{aligned} L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(r_t \mid \delta_t, r_1,\ldots,r_{t-1}\right) &= \frac{1}{(2\boldsymbol{\pi})^{1/2}\sigma_{\delta_t}} \exp\left\{-\frac{(r_t-\mu)^2}{2\sigma_{\delta_t}^2}\right\} \\ L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\delta_t \mid \tilde{\Delta}_t\right) &= m_{\delta_t,\tilde{\Delta}_t} \\ L_{\boldsymbol{\theta},\boldsymbol{\pi}}\left(\tilde{\Delta}_t \mid \tilde{\Delta}_{t-1}\right) &= p_{\tilde{\Delta}_{t-1},\tilde{\Delta}_t} \end{aligned}$$

then

$$Q\left(\boldsymbol{\theta},\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}\right) = -\frac{1}{2}A_1(\hat{\boldsymbol{\theta}}) + A_2(\hat{\mathrm{P}}) + A_3(\tilde{P}) + A_4\left(\tilde{\boldsymbol{\pi}}\right) + Cst$$

where Cst does not depend on the parameters and

$$\begin{aligned} A_1(\hat{\boldsymbol{\theta}}) &= \sum_{t=1}^{n}\sum_{j=1}^{q}\left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log\sigma_{j,t}^2\right) P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\delta_t = j \mid r_1,\ldots,r_n\right\} \\ &= \sum_{t=1}^{n}\sum_{j=1}^{q}\left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log\sigma_{j,t}^2\right) \sum_{i=1}^{d} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_t = \kappa(i,j) \mid r_1,\ldots,r_n\right\}, \\ A_2(\hat{\mathbf{P}}) &= \sum_{i=1}^{q}\sum_{j=1}^{d}\log m_{i,j}\left(\sum_{t=1}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_t = \kappa(j,i) \mid r_1,\ldots,r_n\right\}\right) \\ A_3(\tilde{\boldsymbol{P}}) &= \sum_{i,j=1}^{d}\log p_{i,j}\sum_{t=2}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_{t-1} = i, \tilde{\Delta}_t = j \mid r_1,\ldots,r_n\right\} \\ A_4\left(\tilde{\boldsymbol{\pi}}\right) &= \sum_{i=1}^{d}\log\tilde{\pi}_i P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_1 = i \mid r_1,\ldots,r_n\right\} \end{aligned}$$

Step M: We aim at maximising, with respect to $(\boldsymbol{\theta},\boldsymbol{\pi})$, the estimated log-likelihood $Q\left(\boldsymbol{\theta},\boldsymbol{\pi} \mid \boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}\right)$.

For all t, n, let's denote

$\boldsymbol{\pi}_{t|n} := \left(P\left\{\Delta_t = i \mid r_1,\ldots,r_n\right\}\right)'_{i\le b}$ and $\boldsymbol{\pi}_{t-1,t|n} := \left(P\left\{\Delta_{t-1} = i, \Delta_t = j \mid r_1,\ldots,r_n\right\}\right)'_{i,j\le b}$

We define :

$$\begin{aligned} \bar{\boldsymbol{\theta}} &= \arg\min_{\hat{\boldsymbol{\theta}}\in\Theta} A_1(\hat{\boldsymbol{\theta}}) \\ \bar{\boldsymbol{\theta}} &= \arg\min_{\mu,(\omega_i),(\alpha_i),(\beta_{i,j})} \sum_{t=1}^{n}\sum_{j=1}^{q}\left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log\sigma_{j,t}^2\right)\sum_{i=1}^{d}\boldsymbol{\pi}_{t|n}\left(\kappa(i,j)\right) \end{aligned} \tag{3.10}$$

Using the footnote 10 of Francq and Zakoian (2019), we deduce that maximisation of $A_2(\cdot)$:

$$\begin{aligned}
\bar{m}_{j,i} &= \frac{\sum_{t=1}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_t = \kappa(i,j) \mid r_1,\dots,r_n\right\}}{\sum_{t=1}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_t = i \mid r_1,\dots,r_n\right\}} \\
&= \frac{\sum_{t=1}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_t = \kappa(i,j) \mid r_1,\dots,r_n\right\}}{\sum_{t=1}^{n}\sum_{j=1}^{q} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_t = \kappa(i,j) \mid r_1,\dots,r_n\right\}} \\
\bar{m}_{j,i} &= \frac{\sum_{t=1}^{n} \boldsymbol{\pi}_{t|n}\left(\kappa(i,j)\right)}{\sum_{t=1}^{n}\sum_{j=1}^{q} \boldsymbol{\pi}_{t|n}\left(\kappa(i,j)\right)}
\end{aligned} \tag{3.11}$$

Similarly, $A_3(\cdot)$ yields :

$$\begin{aligned}
\bar{p}_{i,j} &= \frac{\sum_{t=2}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_{t-1} = i, \tilde{\Delta}_t = j \mid r_1,\dots,r_n\right\}}{\sum_{t=2}^{n} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_{t-1} = i \mid r_1,\dots,r_n\right\}} \\
&= \frac{\sum_{t=2}^{n}\sum_{k=1}^{q}\sum_{l=1}^{q} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_{t-1} = \kappa(i,k), \Delta_t = \kappa(j,l) \mid r_1,\dots,r_n\right\}}{\sum_{t=2}^{n}\sum_{l=1}^{q} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_{t-1} = \kappa(i,l) \mid r_1,\dots,r_n\right\}} \\
\bar{p}_{i,j} &= \frac{\sum_{t=2}^{n}\sum_{k,l=1}^{q} \boldsymbol{\pi}_{t-1,t|n}\left(\kappa(i,k),\kappa(j,l)\right)}{\sum_{t=2}^{n}\sum_{l=1}^{q} \boldsymbol{\pi}_{t|n}\left(\kappa(i,l)\right)}
\end{aligned} \tag{3.12}$$

and $A_4(\cdot)$ implies :

$$\begin{aligned}
\bar{\pi}_0(i) &= P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\tilde{\Delta}_1 = i \mid r_1,\dots,r_n\right\} \\
&= \sum_{j=1}^{q} P_{\boldsymbol{\theta}^{(k)},\boldsymbol{\pi}^{(k)}}\left\{\Delta_1 = \kappa(i,j) \mid r_1,\dots,r_n\right\} \\
\bar{\pi}_0(i) &= \sum_{j=1}^{q} \boldsymbol{\pi}_{t|n}\left(\kappa(i,j)\right)
\end{aligned} \tag{3.13}$$

Remark 3.1. *Here, only A_1 needs to be optimized, the other terms can be found analytically once smoothed probabilities $\boldsymbol{\pi}_{t|n}$ and $\boldsymbol{\pi}_{t-1,t|n}$ are known.*

3.5 Computation of Smoothed Probabilities:

The Markov property entails that, given Δ_t, the observations $r_t, r_{t+1}, \dots$ do not convey information on Δ_{t-1}. We hence have

$$\mathbb{P}\left(\Delta_{t-1} = i \mid \Delta_t = j, r_1,\dots,r_n\right) = \mathbb{P}\left(\Delta_{t-1} = i \mid \Delta_t = j, r_1,\dots,r_{t-1}\right)$$

and

$$\boldsymbol{\pi}_{t-1,t|n}(i,j) = \mathbb{P}\left(\Delta_{t-1} = i \mid \Delta_t = j, r_1,\dots,r_n\right)\boldsymbol{\pi}_{t|n}(j) = \frac{p_{i,j}\boldsymbol{\pi}_{t-1|t-1}(i)\boldsymbol{\pi}_{t|n}(j)}{\boldsymbol{\pi}_{t|t-1}(j)}.$$

It remains to compute the smoothed probabilities for $t = n, n-1, \dots, 2$ given by

$$\boldsymbol{\pi}_{t-1|n}(i) = \sum_{j=1}^{b} \boldsymbol{\pi}_{t-1,t|n}(i,j) = \sum_{j=1}^{b} \frac{p_{i,j}\boldsymbol{\pi}_{t-1|t-1}(i)\boldsymbol{\pi}_{t|n}(j)}{\boldsymbol{\pi}_{t|t-1}(j)}$$

3.6 Summary of the Algorithm

Algorithm 1 Summary of the Algorithm

Initialize $\hat{\boldsymbol{\theta}} = (\mu, (\omega_i, \alpha_i)_{i\leq q}, (\beta_{i,j})_{i,j\leq q})$, $(m_{j,i}) = (\mathbb{P}(\delta_1 = j \mid \tilde{\Delta}_1 = i))$ $p_{i,j} = \mathbb{P}(\tilde{\Delta}_2 = j \mid \tilde{\Delta}_1 = i)$, $\tilde{\boldsymbol{\pi}} = \left(\mathbb{P}(\tilde{\Delta}_1 = 1), \ldots, \mathbb{P}(\tilde{\Delta}_1 = d)\right)'$,

while *Not consistent* **do**

Step 1: Compute $\mathbf{P}$ and $\boldsymbol{\pi}_{1|0} = \boldsymbol{\pi}$ by (3.2). For $t = 1$ to n, compute $\boldsymbol{f_{t|t-1}}$ by (3.6) and update:

$$\begin{cases} \boldsymbol{\pi}_{t|t} = \dfrac{\boldsymbol{\pi}_{t|t-1} \odot \boldsymbol{f_{t|t-1}}}{\sum_i \left(\boldsymbol{\pi}_{t|t-1} \odot \boldsymbol{f_{t|t-1}}\right)_i} \\ \boldsymbol{\pi}_{t+1|t} = \mathrm{P}'\boldsymbol{\pi}_{t|t} \end{cases}$$

Step 2: For $t = n$ to 2, compute smoothed probabilities $\boldsymbol{\pi}_{t|n}$ and $\boldsymbol{\pi}_{t-1,t|n}$ using the recurrent formula:

$$\begin{cases} \boldsymbol{\pi}_{t-1|n}(i) = \displaystyle\sum_{j=1}^{b} \frac{p_{i,j}\boldsymbol{\pi}_{t-1|t-1}(i)\boldsymbol{\pi}_{t|n}(j)}{\boldsymbol{\pi}_{t|t-1}(j)} \\ \boldsymbol{\pi}_{t-1,t|n}(i,j) = \dfrac{p_{i,j}\boldsymbol{\pi}_{t-1|t-1}(i)\boldsymbol{\pi}_{t|n}(j)}{\boldsymbol{\pi}_{t|t-1}(j)} \end{cases}$$

Step 3: Using (3.10), (3.11), (3.12) and (3.13), replace the previous values of the parameters by:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \bar{\boldsymbol{\theta}} \\ m_{j,i} &= \bar{m}_{j,i}. \\ p_{i,j} &= \bar{p}_{i,j}, \\ \tilde{\boldsymbol{\pi}} &= \bar{\boldsymbol{\pi}}, \end{aligned}$$

end

Starting from an initial value $\hat{\boldsymbol{\theta}}$, the above equations allow us to obtain a sequence of estimators $\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\pi}^{(k)}\right)_k$ which increase the likelihood. In practice, the sequence converges rapidly to the estimator. We derived in appendix A.4 the gradient to accelerate the minimization of A_1.

3.7 Numerical complexity

For each iteration of the algorithm, knowing that d and q are constants, we can note that the complexity of step 1 and step 2 involving the equations (3.4), (3.6), and (3.7) is $O(n)$. In step 3, the computation of $\bar{m}_{i,j}$, $\bar{p}_{i,j}$, and $\bar{\boldsymbol{\pi}}$ are also in $O(n)$. Besides, the computation of $\bar{\boldsymbol{\theta}}$ has the same complexity as calibrating a classical GARCH model. With GARCH optimization predominant to the others, we can deduce that the overall complexity, with Ni the number of iterations until convergence, is :

$$O(Ni * GARCH) \tag{3.14}$$

In our numerical application, for a sample of approximately 5000 observations, 20 iterations were sufficient to achieve convergence. It is important to notice that, only an update of $\boldsymbol{\theta}$ is needed once a first optimum for the others is found, which reduce the complexity in practice to a simple GARCH. If an important change of regime structure is observed like a new crisis, a full re-calibration will then have to be done.

4 Numerical experiments

In this section, the cross-dependency parameters $(\beta_{ij},\ i \neq j)$ are set to zero. As observed by Haas et al. (2004a), these terms don't lead to significant improvements in practice. Thus, without ambiguity, we rename $\beta = (\beta_{i,i})_i$. Besides, we chose $p = q = 2$ as a good trade-off between efficiency and time-simulation. Finally, we decided to fix μ for both MS-NM-GARCH and GARCH to an empirical value given our case study focus on volatility regardless of the level of this parameter.

4.1 Impact of initialization

The use of the EM algorithm technique requires an initialization. We observed that when dealing with real financial data, it's better to initialize with transition matrix and mixture coefficients having high-value diagonals, with $m_{1,1} > m_{2,2}$, $p_{1,1} > p_{2,2}$; as for the parameters in Table 1. We will refer to these parameters as "type A Parameter". An explanation of these parameter forms can be found in the section below. We also remark in the estimation procedure that it isn't the specific values of the chosen Type A parameters that are important but solely their properties listed above. In Figure 1, we present the value of $-\log L$ with CAC 40 data, from 1990 to 2014, at each iteration of the algorithm with Type A and five other random initializations.

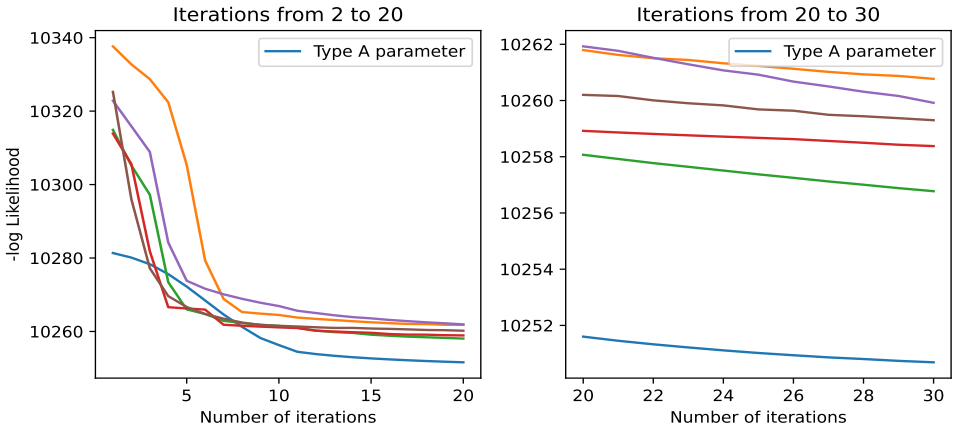

Figure 1 MS(2)-NM(2)-GARCH Loglikelihood with daily CAC40 returns.

4.2 Model parameters after calibration

In Table 1 and 2, we present the fitting parameters on the daily CAC 40 log-returns from 2007 to 2016. We remark that the NM-MS-GARCH acts as a generalization of the GARCH model. Indeed, parameters of the first state are similar to the classical GARCH whereas the other state reflects the crisis period of 2008. Besides, we see that the component volatility $(\sigma_{1,t})$ of financial returns tends to capture persistency and long memory volatilities by having a very high β (> 0.8) and very low ω and α. This form implies a component that is insensitive to the past return and converges almost to an unconditional volatility component. On the other hand, the second component presents relatively high values of ω and α, and a low β (< 0.7). This implies a component with low memory and high sensitivity to variations in the past return, like during crisis. The matrix P defines the persistence of each state and as expected, the system converges to the steady state. Finally the matrix M shows that the steady state acts essentially as a simple GARCH ($m_{1,2}$ negligible) whereas the crisis state acts as a NM-GARCH ($m_{2,1}$ non negligible) with possible excess of Kurtosis.

Remark 4.1. *When $q > 2$ or $d > 2$, the same phenomenon is observed empirically. This modeling allows interpreting $(m_{1,i})_i$ as parameters of persistence and $(m_{q,i})_i$ as crisis factors. The other $(m_{j,i})_i$ smooth the effect with intermediate levels and allow additional degrees of freedom for skew and kurtosis fitting.*

μ	ω	α	β	
0.04	0.03	0.07	0.89	0.00
0.04	1.03	0.44	0.00	0.67

$$M = \begin{pmatrix} 0.93 & 0.07 \\ 0.35 & 0.65 \end{pmatrix} \qquad P = \begin{pmatrix} 0.99 & 0.01 \\ 0.11 & 0.89 \end{pmatrix}$$

Table 1 Parameters estimated from with a MS(2)-NM(2)-GARCH model.

μ	ω	α	β
0.04	0.07	0.11	0.86

Table 2 Parameters estimated with a GARCH(1,1) model.

Remark 4.2. *In this numerical application, we obtained directly a good enough calibration. However, if needed, the model could be constrained to reduce the number of parameters. One simple way is to fix* $m_{1,1} = 1$. *Besides, one should take care to always get values high enough for* $\beta_{1,1}$ *and* α_2 *in order not to reduce the model to a classical GARCH.*

4.3 Return Simulation

Without further researches, the easiest way to simulate first moments for the NM-MS-GARCH model is with Monte Carlo approach.

We first compute $\boldsymbol{\pi}_{t|t} = (\mathbb{P}(\Delta_t = i \mid r_t, \dots, r_1))_i$ using the Hamilton filter and simulate $10,000$ trajectories of Δ_t then r_t (3.1) without re-calibration. We noted in practice that 10,000 trajectories are enough to guarantee at most 5% of error.

4.4 Conditional volatility and density

In Figures 2 and 3, we have plotted the CAC 40 realized volatility curves $(\sqrt{\sum_{k=0}^{T} \boldsymbol{r}_{t-k}^2})_t$, the forecast one $(\mathbb{E}_t\sqrt{\sum_{k=1}^{T+1} \boldsymbol{r}_{t+k}^2})_t$, as well as the Hamilton Filter response from February 2020 to April 2020 ; period corresponding to the COVID-19 impact on financial markets.

We observe with the Hamilton Filter that high probability reflects quite well period of high volatility. The punctual peak in 22.03 is not fully detected though but when we look at the weekly volatility figure, we can see that on average the perturbation is not high enough. Furthermore, NM-MS-GARCH model compared to GARCH one gives advantage only during stress periods, when kurtosis is needed, otherwise we note that both models are equivalent. Finally, we plotted densities in Figure 4 to emphasize our contribution in term of percentile estimation improvement for the class of GARCH models.

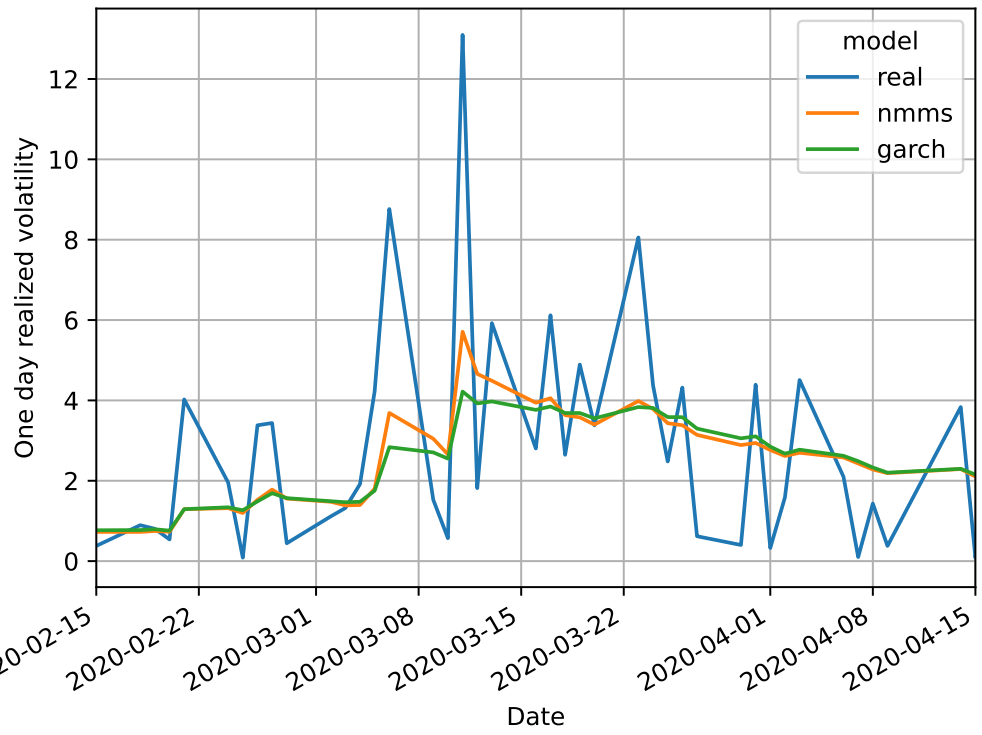

Figure 2 One-day absolute return and forecast of conditional volatility on CAC40 with NM(2)-MS(2)-GARCH and GARCH(1,1) model.

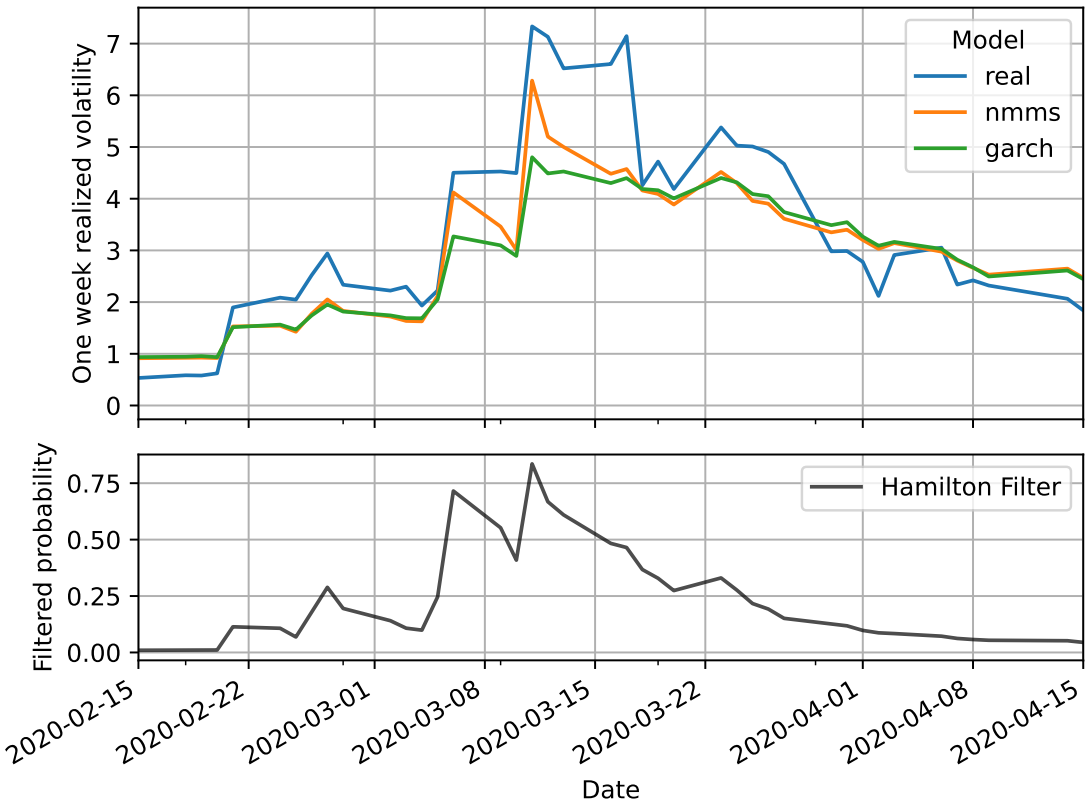

Figure 3 One-week realized volatility and forecast of realized conditional volatility on CAC40 with NM(2)-MS(2)-GARCH and GARCH(1,1) model.

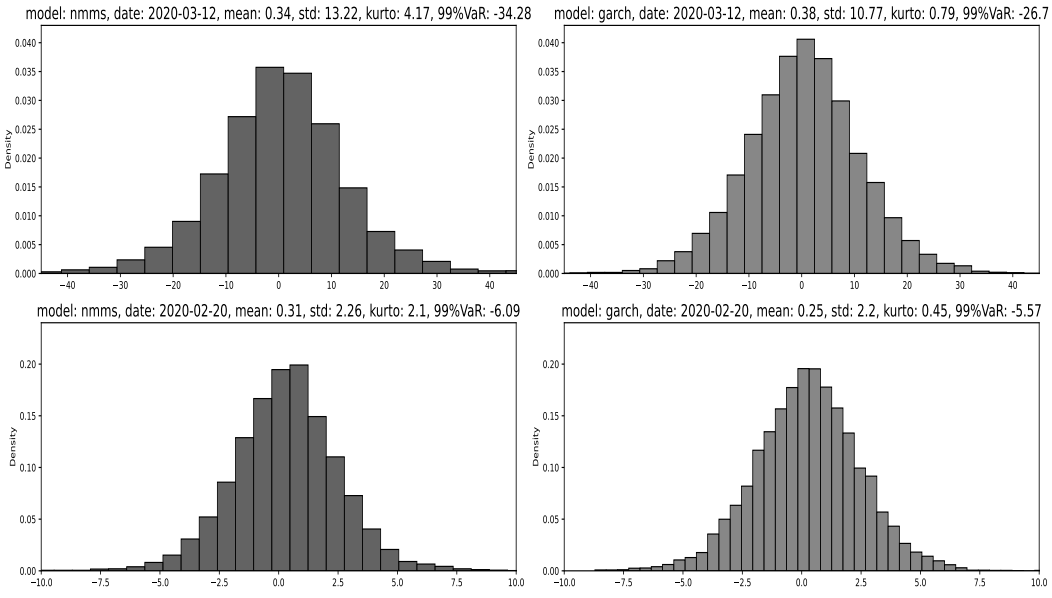

Figure 4 One-week conditional density on CAC40 with NM(2)-MS(2)-GARCH and GARCH(1,1) model in high and low volatility periods

5 Conclusion and further research

In this paper, we introduce a Markov-switching Normal-mixture GARCH model in which the conditional volatility is driven by both a Markov switching sequence and innovations with normal-mixture distributions. We derive the necessary and sufficient conditions for the existence of strictly stationary and second-order stationary solutions. To our knowledge, we have also established, for the first time, the strong consistency of the maximum likelihood estimator for a Markov switching model with a pure GARCH-type component.

We apply this model on a data set of CAC 40 returns and show that this new model gives better performance during crisis period for a similar level of calibration time. Indeed, it is well known that simple GARCH models cannot generate and maintain for too long high level of volatility due to their high persistence factor, entailing an under estimated risk in portfolio allocation. In our study, we show how to mitigate these issues without additional complexity. On one hand, the NM feature can create excess of kurtosis implying higher value of extreme percentiles. On the other hand, the MS feature can produce long period of high volatility if observed during the calibration and amplify the NM's effect. Thus, this new model makes a better monitoring of risk in tactical portfolio optimization and allows the use of indicators based on percentile like Value At Risk.

Besides, we derived a simple algorithm to calibrate the model and show that once the transition matrices are fitted, the complexity is equivalent to a simple GARCH model. The overall complexity is empirically less than 20 times a GARCH one. In our study case, we obtained a good calibration quite quickly. However, if needed, one can constraint the model parameters to accelerate the calibration.

Finally, a topic left to be addressed is the generation of first moments (mean, variance, skew and kurtosis) and percentiles without use of Monte Carlo. Nevertheless, the natural extension to correlated assets, trivial with Monte Carlo, may add significant complications to another approach.

References

Carol Alexander and Emese Lazar. Normal mixture garch (1, 1): Applications to exchange rate modelling. *Journal of Applied Econometrics*, 21(3):307–336, 2006.

István Berkes, Lajos Horváth, and Piotr Kokoszka. GARCH processes: structure and estimation. *Bernoulli*, 9(2):201 – 227, 2003. doi: 10.3150/bj/1068128975. URL https://doi.org/10.3150/bj/1068128975.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3): 307–327, 1986.

Jun Cai. A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, 12(3): 309–316, 1994.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Randal Douc, Eric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. 2004.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*, pages 987–1007, 1982.

Robert F Engle and Jose Gonzalo Rangel. The spline-garch model for low-frequency volatility and its global macroeconomic causes. *The review of financial studies*, 21(3):1187–1222, 2008.

Robert F Engle, Eric Ghysels, and Bumjean Sohn. Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 95(3):776–797, 2013.

Christian Francq and Jean-Michel Zakoian. Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli*, 10(4):605–637, 2004.

Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

Christian Francq, Michel Roussignol, and Jean-Michel Zakoian. Conditional heteroskedasticity driven by hidden markov chains. *Journal of Time Series Analysis*, 22(2):197–220, 2001.

Markus Haas, Stefan Mittnik, and Marc S Paolella. Mixed normal conditional heteroskedasticity. *Journal of financial Econometrics*, 2(2):211–250, 2004a.

Markus Haas, Stefan Mittnik, and Marc S Paolella. A new approach to markov-switching garch models. *Journal of financial Econometrics*, 2(4):493–530, 2004b.

James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, 64(1-2):307–333, 1994.

Ken Johnston and Elton Scott. Garch models and the stochastic process underlying exchange rate price changes. *Journal of Financial and Strategic Decisions*, 13(2):13–24, 2000.

Baye Matar Kandji. On the growth rate of superadditive processes and the stability of functional garch models. 2024.

J. F. C. Kingman. Subadditive Ergodic Theory. *The Annals of Probability*, 1:883 – 899, 1973.

Gary GJ Lee and Robert F Engle. A permanent and transitory component model of stock return volatility. *Available at SSRN 5848*, 1993.

Taewook Lee and Sangyeol Lee. Normal mixture quasi-maximum likelihood estimator for garch models. *Scandinavian Journal of Statistics*, 36(1):157–170, 2009.

Daniel B Nelson. Asymptotic filtering theory for multivariate arch models. *Journal of Econometrics*, 71 (1-2):1–47, 1996.

Peter M Robinson. Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, 47(1):67–84, 1991.

Jean-Michel Zakoian. Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, 18 (5):931–955, 1994.

A Appendix: Complementary Proofs

A.1 A decomposition lemma for random variables

Lemma A.1. *Let $\mathcal{S} = (\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let $\boldsymbol{X}$ be a real-valued random variable, and $\boldsymbol{\eta}$ be a $\mathbb{R}^n$ random variable. Let f be a measurable function from $\mathbb{R}^n$ to $\mathbb{R}$. If $\boldsymbol{X}$ and $f(\boldsymbol{\eta})$ have the same distribution, then: If $\mathcal{S}$ sufficiently rich (we can enrich it if necessary), then there exists a random variable $\boldsymbol{\eta}'$ with the distribution of $\boldsymbol{\eta}$ such that $\boldsymbol{X} = f(\boldsymbol{\eta}')$ a.s.*

Proof. Let $\kappa : \mathbb{R} \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ be the regular conditional probability distribution of $\boldsymbol{\eta}$ given $f(\boldsymbol{\eta})$. Let $U_1, U_2, \cdots, U_n$ n iid uniform distribution on $[0,1]$, independent of $\boldsymbol{X}$ and $\boldsymbol{\eta}$ [2]. By standard construction [3], for all $f(\boldsymbol{\eta})^\omega$ there exists a measurable function $g(f(\boldsymbol{\eta})^\omega, \cdot)$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ such that $g(f(\boldsymbol{\eta}), (U_i)_{i\leq n}) \mid f(\boldsymbol{\eta})$ has $\kappa(f(\boldsymbol{\eta}), \cdot)$ as (conditional) distribution. It follows by the definition of κ that $f\left(g(f(\boldsymbol{\eta}), (U_i)_{i\leq n})\right) = f(\boldsymbol{\eta})$ a.s. Let $\boldsymbol{\eta}' = g(\boldsymbol{X}, (U_i)_{i\leq n})$. Since $(\boldsymbol{X}, (U_i)_{i\leq n}))$ and $(f(\boldsymbol{\eta}), (U_i)_{i\leq n}))$ have the same distribution then $f(\boldsymbol{\eta}') = \boldsymbol{X}$ a.s. and

Now it is reminded to show that $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$ have the same distribution. The definition of $\boldsymbol{\eta}'$ implies that the regular conditional probability distribution of $\boldsymbol{\eta}'$ given $\boldsymbol{X}$ is also κ. For all $A \in \mathcal{B}(\mathbb{R}^n)$ On has

$$\mathbb{P}(\boldsymbol{\eta} \in A) = \int_{\mathbb{R}} \kappa(x, A) \mathbb{P}_{f(\boldsymbol{\eta})}(dx) = \int_{\mathbb{R}} \kappa(x, A) \mathbb{P}_{\boldsymbol{X}}(dx) = \mathbb{P}(\boldsymbol{\eta}' \in A),$$

because $f(\boldsymbol{\eta})$ and $\boldsymbol{X}$ have the same distribution. This concludes the proof. □

Remark A.1.

1. *The condition that $\boldsymbol{X}$ and $f(\boldsymbol{\eta})$ are defined in the same probability space is not restrictive. One can consider that that is not the case, and adapt the proof by considering $(U_i)_{i\leq n}$ independent of $f(\boldsymbol{\eta})$ in the space related to $f(\boldsymbol{\eta})$ to define g, by extending the space without restriction if necessary; and also define another iid uniform distribution on $[0,1]$ couple $(U'_i)_{i\leq n}$ independent of $f(\boldsymbol{X})$ in $\mathcal{S}$ to define $\boldsymbol{\eta}'$, if the space is sufficiently rich or by extending if necessary.*

2. *If one can define different n-couples of iid uniform distribution independent of $\boldsymbol{X}$ in $\mathcal{S}$, then it is clear that our construction is not unique. The uniqueness also depend on the structure f in the support of $\boldsymbol{\eta}$. Indeed, one can remark that if f is bijective in that support, then $\kappa(x, A) = \delta_{f^{-1}(x)}(A)$ which is the distribution of the constant random variable $f^{-1}(x)$ for fixed x. Thus g doesn't depend on $(U_i)_{i\leq n}$ and is equal to f^{-1}.*

A.2 Stationarity

Let us start with some conventions. For all reel matrices $A = (a_{ij})$ and B we say that $A > 0$ (resp. $A \geq 0$) if for all i, j $a_{ij} > 0$ (resp. $a_{ij} \geq 0$); $A > B$ if $A - B > 0$.

Proof. We first show that if $\rho(\boldsymbol{Q}) < 1$ then $\boldsymbol{\gamma} < 0$. For all all matrix $A = (a_{ij})$, $\|A\| = \sum_{ij} |a_{ij}|$. Since $B_t^{(k)}$ has positive elements, for all k, we have

$$\begin{aligned}\mathbb{E}\|B_t^{(k)}\| = \|\mathbb{E}B_t^{(k)}\mathbf{1}'_{(d)}\| = \|\mathbb{E}\{\mathbb{E}B_t^{(k)}\mathbf{1}'_{(d)}\} \mid (\tilde{\Delta}_t)_{t\geq t-k}\| = \|\mathbb{E}\{\tilde{B}(\tilde{\Delta}_t)\cdots\tilde{B}(\tilde{\Delta}_{t-k+1})\mathbf{1}'_{(d)}\}\| \\ = \left\|\mathbf{I}'\boldsymbol{Q}^k\mathbf{J}'\right\| = \mathbf{1}'_{(dq)}\boldsymbol{Q}^k\mathbf{J}'\end{aligned}$$

where $\mathbf{I}' = (I_q, \ldots, I_q)$ is $q \times dq$ matrix and $\mathbf{J} = (\tilde{\pi}_1\mathbf{1}'_{(d)}, \ldots, \pi_d\mathbf{1}'_{(d)})$. Since $\rho(\boldsymbol{Q}) < 1$ implies that $\lim_{k\to+\infty} \frac{1}{k}\|\boldsymbol{Q}^k\| = \log\rho(\boldsymbol{Q}) < 0$, it follows by the Jensen inequality that $\boldsymbol{\gamma} \leq \lim_{n\to+\infty} \frac{1}{k}\log\{\mathbb{E}\|B_t^{(k)}\|\} < 0$. Thus the existence of a (unique) strictly stationary (and ergodic) solution follows from Theorem 2.1. Let us now compute the variance. By the total variance formula,

$$\operatorname{var}(r_t) = \mathbb{E}\{\operatorname{var}\left(r_t \mid (\tilde{\Delta}_s)_{s\leq t}\right)\} + \operatorname{var}\left(\mathbb{E}\{r_t \mid (\tilde{\Delta}_s)_{s\leq t}\}\right) = \mathbb{E}\boldsymbol{\epsilon}_t^2 + \operatorname{var}\mu_t \tag{A.1}$$

[2] The possibility to define these random variables depends on the richness of $\mathcal{S}$, however, one can always use a product construction to enlarge the probability space if necessary.

[3] Using the multiplication rule for multivariate distributions and the inverse transformation method.

For the first term, by argument already used we have

$$\mathbb{E}\boldsymbol{\epsilon}_t^2 = \mathbb{E}(\sum_{k=1}^{q} \mathbb{1}_{\boldsymbol{\delta}_t=k}\sigma_{t,k}^2) = \sum_{k=0}^{\infty} \mathbb{E}\{m(\tilde{\Delta}_t)\tilde{B}(\tilde{\Delta}_{t-1})\cdots\tilde{B}(\tilde{\Delta}_{t-k})\omega\}$$
$$= \sum_{k=0}^{\infty} \mathbf{1}'_{(d)}\psi\boldsymbol{Q}^k\boldsymbol{z} = \mathbf{1}'_{(d)}\psi\left(I_{dq}-\boldsymbol{Q}\right)^{-1}\boldsymbol{z}$$

The second part of Eq. (A.1) is easy to compute, this prove the first part of the theorem.

Now suppose that (r_t) is a positive second order stationary solution of (2.1). Let $\bar{P} = \sum_{k=0}^{\infty}\boldsymbol{Q}^k$. Since $\mathbb{E}r_t \geq \sum_{k=0}^{\infty}\mathbb{E}\{m(\tilde{\Delta}_t)\tilde{B}(\tilde{\Delta}_{t-1})\cdots\tilde{B}(\tilde{\Delta}_{t-k})\omega\} = \sum_{k=0}^{\infty}\mathbf{1}'_{(d)}\psi\boldsymbol{Q}^k\boldsymbol{z}$, thus this latter sum is finite. It follows that its reminder $(\mathbf{1}'_{(d)}\psi\boldsymbol{Q}^k\bar{P}\boldsymbol{z})_k$ converges to zero. Assumption (i) implies that $\bar{P} > 0$, and by Assumption (ii), $\sup_i \boldsymbol{z}_i > 0$. Thus $\bar{P}\boldsymbol{z} > 0$. It follows that

$$(\mathbf{1}'_{(d)}\psi\boldsymbol{Q}^k)_k \text{ converges to zero.} \tag{A.2}$$

Assumption (i) also implies that for all $i, j \leq dq$ there exists an integer r such that the i,j entry, $\boldsymbol{Q}^r_{ij}$, of $\boldsymbol{Q}^r$ is strictly positive. Since $\mathbf{1}'_{(d)}\boldsymbol{\psi}$ has a strictly positive entry, then for all $i \leq dq$ there exist r_i such that the entry i of $\mathbf{1}'_{(d)}\boldsymbol{\psi}\boldsymbol{Q}^{r_i}$ is strictly positive. Thus $\mathbf{1}'_{(d)}\boldsymbol{\psi}\boldsymbol{Q}^r > 0$. Rewriting $\mathbf{1}'_{(d)}\psi\boldsymbol{Q}^k$ by $\mathbf{1}'_{(d)}\psi\boldsymbol{Q}^{r_i}\boldsymbol{Q}^{k-r_i}$ for all i, it follows from Eq. (A.2) that the row i of $\boldsymbol{Q}^k$ converges to 0 for all $i \leq dq$. Therefore, $\boldsymbol{Q}^k$ converges to 0, thus Fekete's lemma implies that $\rho(\boldsymbol{Q}) < 1$. □

A.3 Consistency of the Maximum likelihood estimator

The parameter space Θ is compact, compatible with the following conditions: for all $\theta \in \Theta$, the associated Markov chain $(\tilde{\Delta}_t)$ is irreducible and aperiodic, $U > 0$, $\min_i \omega_i > 0$, and $\rho(\boldsymbol{\beta}) < 1$. It is assumed that the true parameter value θ_0 belongs to Θ, and that $\rho(\boldsymbol{Q}(\theta_0)) < 1$.

Let $\pi_{t|\cdot}(i) = \mathbb{P}(\Delta_t = i \mid \boldsymbol{r}_{t-1}, \boldsymbol{r}_{t-2}, \ldots)$, $g_{m,\theta}(\cdot \mid \boldsymbol{r}_{t-m}, \boldsymbol{r}_{t-m-1}, \ldots)$ be the density function of $(\boldsymbol{r}_t, \ldots, \boldsymbol{r}_{t-m+1})$ given the σ-field generated by $\boldsymbol{r}_{t-m}, \boldsymbol{r}_{t-m-1}, \ldots$, and $f_{\theta,t,i}(\cdot)$ be the density function of $\boldsymbol{r}_t$ given the σ-field generated by $\Delta_t = i, \boldsymbol{r}_{t-1}, \boldsymbol{r}_{t-2}, \ldots$. Without ambiguity, define $g_\theta\left(r_t \mid r_{t-1}, r_{t-2}, \ldots\right) := g_{1,\theta}\left(r_t \mid r_{t-1}, r_{t-2}, \ldots\right)$

The following identifiability condition is needed for the strong consistency of $(\widehat{\boldsymbol{\theta}}_n)$.

Assumption A: For all $\theta \in \Theta$, if $g_\theta\left(r_t \mid r_{t-1}, r_{t-2}, \ldots\right) = g_{\theta_0}\left(r_t \mid r_{t-1}, r_{t-2}, \ldots\right)$ $P_{\theta_0}-$ a.s, then $\theta = \theta_0$.

Theorem A.1. *Under **Assumption A**, $(\widehat{\boldsymbol{\theta}}_n)$ converges almost surely to θ_0 as $n \to \infty$.*

The proof of the consistency of the maximum likelihood estimator relies on the following lemmas.

Lemma A.2. *For all i, we have*

$$\inf_{\theta\in\Theta} \pi_{t|\cdot}(i) > 0 \quad P_{\theta_0}\text{-a.s.}$$

Proof. By the same argument used to establish (3.7), we have

$$\pi_{t|\cdot} = \mathrm{P}'\frac{f_{\theta,t-1}(r_{t-1})}{\mathbf{1}'\left\{f_{\theta,t-1}(r_{t-1}) \odot \pi_{t-1|\cdot}\right\}} \odot \pi_{t-1|\cdot} \geq \mathrm{P}_{t-1}\pi_{t-1|\cdot}, \tag{A.3}$$

where $\mathrm{P}_t(i,j) = \mathrm{P}'_{i,j}\frac{f_{\theta,t-1,j}(r_{t-1})}{(1/b)\max_k\{f_{\theta,t-1,k}(r_{t-1})\}}$, because $\max_i \pi_{t-1|\cdot}(i) > 1/b$.

The irreducibility and aperiodicity assumptions, together with the condition $U > 0$, imply that (Δ_t) is primitive; i.e., for all θ', there exists a strictly positive integer k such that $(\mathrm{P}')^k(\theta') > 0$. Since for all j and θ, $f_{\theta,t-1,j}(\cdot)$ are strictly positive functions, it follows that

$$\prod_{l=1}^{k} \mathrm{P}_{t-1-l}(\theta') > 0.$$

From (A.3), we have

$$\pi_{t|\cdot} \geq \left(\prod_{l=1}^{k} \mathrm{P}_{t-1-l} \right) \pi_{t-1-k|\cdot}.$$

Since $\max_i \pi_{t-1-k|\cdot}(i) > 1/b$, and the components of P_t are continuous in θ and strictly positive at θ', it follows that for all θ' there exists a continuous function h (depending on θ') such that $\pi_{t|\cdot} \geq h$ and $h(\theta') > 0$. By continuity, there exists an open ball $V_{\theta'}$ centered at θ' such that $\inf_{\theta \in V_{\theta'} \cap \Theta} h(\theta) > 0$, and hence $\inf_{\theta \in V_{\theta'} \cap \Theta} \pi_{t|\cdot} > 0$.

By compactness of Θ, there exists a finite sub-cover of the form $V_{\theta_1}, V_{\theta_2}, \ldots, V_{\theta_M}$. It follows that

$$\inf_{\theta \in \Theta} \pi_{t|\cdot} \geq \min_{l \leq M} \inf_{\theta \in V_{\theta_l} \cap \Theta} \pi_{t|\cdot} > 0.$$

□

Lemma A.3. *Let $\tilde{L}_\theta(\cdot)$ be the density of $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ given the σ-field generated by $\boldsymbol{r}_0, \boldsymbol{r}_{-1}, \ldots$,*

$$\lim_{n \to \infty} \frac{1}{n} \sup_{\theta \in \Theta} \left| \log \frac{\tilde{L}_\theta(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)}{L_\theta(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)} \right| = 0 \quad P_{\theta_0} - a.s$$

Proof. Let us first rewrite the likelihood in the Matrix product formula introduced by Francq et al. (2001).

Let $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^b$, $\pi_0 = (\pi_0(1), \ldots, \pi_0(b))' \in \mathbb{R}^b$ and

$$M_\theta(r_t, \ldots, r_1) = \begin{pmatrix} p_{1,1}\phi_1(r_t, \ldots, r_1) & p_{2,1}\phi_1(r_t, \ldots, r_1) & \cdots & p_{b,1}\phi_1(r_t, \ldots, r_1) \\ p_{1,2}\phi_2(r_t, \ldots, r_1) & p_{2,2}\phi_2(r_t, \ldots, r_1) & \cdots & p_{b,2}\phi_2(r_t, \ldots, r_1) \\ \vdots & \vdots & & \vdots \\ p_{1,b}\phi_b(r_t, \ldots, r_1) & p_{2,b}\phi_b(r_t, \ldots, r_1) & \cdots & p_{b,b}\phi_b(r_t, \ldots, r_1) \end{pmatrix}.$$

One can check that,

$$L_\theta(r_1, \ldots, r_n) = \mathbf{1}' \left\{ \prod_{k=1}^{n} M_\theta(r_n, \ldots, r_1) \right\} \pi_0. \tag{A.4}$$

The expression of $\tilde{L}_\theta(r_1, \ldots, r_n)$ is similar to that of $L_\theta(r_1, \ldots, r_n)$ in (A.4). It is obtained from $L_\theta(r_1, \ldots, r_n)$ by replacing $\pi_0(i)$ with $\pi_{0|\cdot}(i)$ and $\phi_i(r_t, \ldots, r_1)$ with $\tilde{\phi}_i(r_t, \ldots, r_1, r_0, \ldots)$, where $\tilde{\phi}_i$ is defined by substituting $(\boldsymbol{\sigma}_t)$ with $(\tilde{\boldsymbol{\sigma}}_t)$, the strictly stationary, ergodic, and non-anticipative solution of (3.4), in the expression of $\phi_i(r_t, \ldots, r_1)$.

Let us define $\bar{L}_\theta(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ by substituting π_0 with $\pi_{0|\cdot}$ in the expression of $L_\theta(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ in (A.4). We have

$$\frac{1}{n} \sup_{\theta \in \Theta} \left| \log \frac{\tilde{L}_\theta(r_1, \ldots, r_n)}{L_\theta(r_1, \ldots, r_n)} \right| \leq \frac{1}{n} \sup_{\theta \in \Theta} \left| \log \frac{\tilde{L}_\theta(r_1, \ldots, r_n)}{\bar{L}_\theta(r_1, \ldots, r_n)} \right| + \frac{1}{n} \sup_{\theta \in \Theta} \left| \log \frac{\bar{L}_\theta(r_1, \ldots, r_n)}{L_\theta(r_1, \ldots, r_n)} \right|.$$

The goal is to prove that the two terms on the right-hand side converge to 0.

$$\begin{aligned} \sup_{\theta \in \Theta} \left| \log \frac{\bar{L}_\theta(r_1, \ldots, r_n)}{L_\theta(r_1, \ldots, r_n)} \right| &\leq \sup_{\theta \in \Theta} \max_{i \leq q} \left| \log \frac{\pi_0(i)}{\pi_{0|\cdot}(i)} \right| \\ &\leq \sup_{\theta \in \Theta} \max_{i \leq q} \left(|\log \pi_0(i)| + \left| \log \pi_{0|\cdot}(i) \right| \right). \end{aligned}$$

By the compactness of Θ, Lemma A.2, and the condition $\pi_0(i) > 0$ for all i, the terms on the right-hand side of the inequality above are finite. Since this upper bound does not depend on n, the result follows by dividing by n and taking the limit.

Now let us show that

$$\lim_{n\to\infty}\frac{1}{n}\sup_{\theta\in\Theta}\left|\log\frac{\tilde{L}_\theta(r_1,\ldots,r_n)}{\overline{L}_\theta(r_1,\ldots,r_n)}\right|=0\quad P_{\theta_0}\text{-a.s.}.$$

We have

$$\begin{aligned}\sup_{\theta\in\Theta}\frac{1}{n}\left|\log\frac{\tilde{L}_\theta(r_1,\ldots,r_n)}{\overline{L}_\theta(r_1,\ldots,r_n)}\right|&\leq\sup_{\theta\in\Theta}\frac{1}{n}\sum_{t=1}^{n}\max_j\left|\log\tilde{\phi}_j(r_t,\ldots)-\log\phi_j(r_t,\ldots,r_1)\right|\\&\leq\frac{1}{2n}\sum_{t=1}^{n}\sup_{\theta\in\Theta}\left((r_t-\mu)^2\max_j\left|\frac{1}{\sigma^2_{j,t}}-\frac{1}{\tilde{\sigma}^2_{j,t}}\right|+\max_j\left|\log\frac{\tilde{\sigma}^2_{j,t}}{\sigma^2_{j,t}}\right|\right)\end{aligned}$$

For all t, we have $\tilde{\boldsymbol{\sigma}}_t^2-\boldsymbol{\sigma}_t^2=\boldsymbol{B}^t(\tilde{\boldsymbol{\sigma}}_0^2-\boldsymbol{\sigma}_0^2)$. Applying the inequality $|\log(x/y)|\leq\frac{|x-y|}{x\vee y}$ for $x,y>0$, we deduce

$$\sup_{\theta\in\Theta}\frac{1}{n}\left|\log\frac{\tilde{L}_\theta(r_1,\ldots,r_n)}{\overline{L}_\theta(r_1,\ldots,r_n)}\right|\leq\frac{K}{n}\sum_{t=1}^{n}\left[(r_t-\mu)^2+1\right]\rho^t,$$

where K is an $\mathcal{F}_0$-measurable random variable (independent of t) and $\rho=\sup_{\theta\in\Theta}\rho(\boldsymbol{B})<1$ due to the compactness of Θ. Here, ρ denotes the spectral radius function.

It is well known that for a stationary sequence $(\boldsymbol{x}_n)$ such that $\mathbb{E}|\boldsymbol{x}_0|^s<\infty$ for some $s>0$, we have $\rho^n\boldsymbol{x}_n\to0$ as $n\to\infty$ (see Francq and Zakoian (2004, Proof of Theorem 2.1)). Thus, since $\mathbb{E}r_0^2<\infty$, the result follows the Cesàro lemma. □

Lemma A.4. *For all $\theta_1\in\Theta$, different from θ_0, there exists a neighborhood $V(\theta_1)$ of θ such that*

$$\limsup_{n\to\infty}\frac{1}{n}\sup_{\theta\in V(\theta_1)}\log\frac{\tilde{L}_\theta(r_1,\ldots,r_n)}{\tilde{L}_{\theta_0}(r_1,\ldots,r_n)}<0\quad P_{\theta_0}\text{-a.s.}$$

Proof. To prove this lemma, we use the expression of the log-likelihood in (A.5), where $\tilde{L}_\theta$ is obtained by initializing the filter (3.7) with $\pi_{0|\cdot}(i)$ and $(\tilde{\boldsymbol{\sigma}}_t)$ is used instead of $(\boldsymbol{\sigma}_t)$ to compute the conditional density evaluated at r_t: $f_{\theta,t,i}(r_t)$, as in the expression (3.5).

We then have

$$\tilde{L}_\theta=\sum_{t=1}^{n}\log g_\theta\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)\tag{A.5}$$

where

$$g_\theta\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)=\sum_{i=1}^{b}\pi_{t|\cdot}(i)f_{\theta,t,i}(r_t).$$

For any $\theta\in\Theta$ and any positive integer k, let $V_k(\theta)$ be the open ball of center θ and radius $1/k$. Since $\mathbb{E}r_0^2<\infty$, it is easy to show that

$$\mathbb{E}\sup_{\theta\in\Theta}\log\left|g_\theta\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)\right|\leq\infty.$$

Thus, by the ergodic theorem followed by the dominated convergence theorem, we have

$$\begin{aligned}\lim_{k\to\infty}\limsup_{n\to\infty}\frac{1}{n}\sup_{\theta\in V_k(\theta_1)}\log\frac{\tilde{L}_\theta\left(r_1,\ldots,r_n\right)}{\tilde{L}_{\theta_0}\left(r_1,\ldots,r_n\right)}&=\lim_{k\to\infty}\mathbb{E}\sup_{\theta\in V_k(\theta_1)}\log\frac{g_\theta\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}{g_{\theta_0}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}\quad P_{\theta_0}\text{-a.s.}\\&=\mathbb{E}\log\frac{g_{\theta_1}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}{g_{\theta_0}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}.\end{aligned}$$

By Jensen's inequality,

$$\mathbb{E}_{\theta_0}\log\frac{g_{\theta_1}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}{g_{\theta_0}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}<\log\mathbb{E}_{\theta_0}\frac{g_{\theta_1}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}{g_{\theta_0}\left(r_t\mid r_{t-1},r_{t-2},\ldots\right)}=0.$$

This concludes the proof. □

Proof of Theorem A.1. The result follows from Lemma A.4, Lemma A.3, and the compactness of Θ. □

A.4 The Gradient computation

Recall that only A_1 depends on $\hat{\theta}$, the parameter related to the components' volatilities.

$\frac{\partial A_1}{\partial \hat{\theta}_i} = \sum_{t=1}^{n} \sum_{j=1}^{q} P_{\theta^{(k)},\pi^{(k)}} \{\boldsymbol{\delta}_t = j \mid r_1, \ldots, r_n\} \frac{\partial}{\partial \hat{\theta}_i} \left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log \sigma_{j,t}^2 \right)$

One has:

- $\frac{\partial}{\partial \mu} \left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log \sigma_{j,t}^2 \right) = -\frac{2(r_t-\mu)}{\sigma_{j,t}^2} - \frac{(r_t-\mu)^2 \frac{\partial}{\partial \mu}\sigma_{j,t}^2}{\sigma_{j,t}^4} + \frac{\frac{\partial}{\partial \mu}\sigma_{j,t}^2}{\sigma_{j,t}^2}$ where $(\frac{\partial}{\partial \mu}\sigma_{j,t})_j$ are defined recursively by

$$\frac{\partial}{\partial \mu}\sigma_t^2 = \frac{\partial}{\partial \mu}\boldsymbol{c}_t + B\frac{\partial}{\partial \mu}\sigma_{t-1}^2, \text{ and } \frac{\partial}{\partial \mu}\boldsymbol{c}_t = (-2\alpha_i(\boldsymbol{r}_{t-1} - \mu))_i.$$

- $\frac{\partial}{\partial \omega_i} \left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log \sigma_{j,t}^2 \right) = -\frac{(r_t-\mu)^2 \frac{\partial}{\partial \omega_i}\sigma_{j,t}^2}{\sigma_{j,t}^4} + \frac{\frac{\partial}{\partial \omega_i}\sigma_{j,t}^2}{\sigma_{j,t}^2}$ where $(\frac{\partial}{\partial \omega_i}\sigma_{j,t})_j$ are defined recursively by

$$\frac{\partial}{\partial \omega_i}\sigma_t^2 = \frac{\partial}{\partial \omega_i}\boldsymbol{c}_t + B\frac{\partial}{\partial \omega_i}\sigma_{t-1}^2, \text{ and } (\frac{\partial}{\partial \omega_i}\boldsymbol{c}_t)_i = 1, \ (\frac{\partial}{\partial \omega_i}\boldsymbol{c}_t)_j = 0 \text{ if } j \neq i.$$

- $\frac{\partial}{\partial \alpha_i} \left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log \sigma_{j,t}^2 \right) = -\frac{(r_t-\mu)^2 \frac{\partial}{\partial \alpha_i}\sigma_{j,t}^2}{\sigma_{j,t}^4} + \frac{\frac{\partial}{\partial \alpha_i}\sigma_{j,t}^2}{\sigma_{j,t}^2}$ where $(\frac{\partial}{\partial \alpha_i}\sigma_{j,t})_j$ are defined recursively by

$$\frac{\partial}{\partial \alpha_i}\sigma_t^2 = \frac{\partial}{\partial \alpha_i}\boldsymbol{c}_t + B\frac{\partial}{\partial \alpha_i}\sigma_{t-1}^2, \text{ and } (\frac{\partial}{\partial \alpha_i}\boldsymbol{c}_t)_i = (r_{t-1} - \mu)^2, \ (\frac{\partial}{\partial \alpha_i}\boldsymbol{c}_t)_j = 0 \text{ if } j \neq i.$$

- $\frac{\partial}{\partial \beta_{i,k}} \left(\frac{(r_t-\mu)^2}{\sigma_{j,t}^2} + \log \sigma_{j,t}^2 \right) = -\frac{(r_t-\mu)^2 \frac{\partial}{\partial \beta_{i,k}}\sigma_{j,t}^2}{\sigma_{j,t}^4} + \frac{\frac{\partial}{\partial \beta_{i,k}}\sigma_{j,t}^2}{\sigma_{j,t}^2}$ where $(\frac{\partial}{\partial \beta_{i,k}}\sigma_{j,t})_j$ are defined recursively by

$$\frac{\partial}{\partial \beta_{i,k}}\sigma_t^2 = B^{(i,k)}\sigma_{t-1}^2 + B\frac{\partial}{\partial \beta_{i,k}}\sigma_{t-1}^2,$$

where $B^{(i,k)}$ is a $q \times q$ matrix with (i,k) th element 1 and all other elements 0.