A new approach to volatility modeling: the factorial hidden Markov volatility model

Maciej Augustyniak[a,b], Luc Bauwens[c] and Arnaud Dufays[*,d,e]

[a] *Département de mathématiques et de statistique, Université de Montréal, Canada*
[b] *Quantact Laboratory, Centre de recherches mathématiques, Canada*
[c] *CORE, Université catholique de Louvain, Belgium*
[d] *Département d'économique, Université Laval, Canada*
[e] *CRREP research associate*

Version: April 30, 2018

Abstract

A new process — the factorial hidden Markov volatility (FHMV) model — is proposed to model financial returns or realized variances. Its dynamics are driven by a latent volatility process specified as a product of three components: a Markov chain controlling volatility persistence, an independent discrete process capable of generating jumps in the volatility, and a predictable (data-driven) process capturing the leverage effect. An economic interpretation is attached to each one of these components. Moreover, the Markov chain and jump components allow volatility to switch abruptly between thousands of states, and the transition matrix of the model is structured to generate a high degree of volatility persistence. An empirical study on six financial time series shows that the FHMV process compares favorably to state-of-the-art volatility models in terms of in-sample fit and out-of-sample forecasting performance over time horizons ranging from one to one hundred days.

JEL classification: C22, C51, C58

Keywords: volatility, Markov-switching, hierarchical hidden Markov model, persistence, leverage effect

Maciej Augustyniak acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada. Arnaud Dufays acknowledges financial support from the Fonds de recherche du Québec – Société et culture and from the Social Sciences and Humanities Research Council of Canada via the project 430-2017-00215. The authors are grateful to three anonymous referees for providing valuable comments on an earlier draft as well as to seminar participants at University of Namur and to 2017 SoFiE conference participants at the NYU Stern School of Business (in particular to Eric Ghysels who was the paper's discussant).

[*] Corresponding author.
Email addresses: augusty@dms.umontreal.ca (Maciej Augustyniak), luc.bauwens@uclouvain.be (Luc Bauwens), arnaud.dufays@ecn.ulaval.ca (Arnaud Dufays).

1 Introduction

Building on the seminal contribution of Goldfeld and Quandt (1973), Hamilton (1989) popularized the use of regime-switching models in economics and finance. These models allow us to model sharp changes in the dynamics of economic or financial time series by introducing a finite-valued latent stochastic process that governs the evolution of the parameters of the time series model. In most applications this latent process is a Markov chain and, consequently, Markov-switching and hidden Markov models are sometimes used interchangeably with regime-switching models. In the past twenty-five years, the emphasis in the literature has been on models with a relatively low number of states — between two and four (e.g., Ang and Bekaert, 2002; Bauwens et al., 2014; Dai et al., 2007). On one hand, this choice is motivated by parsimony because the number of parameters in the transition matrix of the Markov chain increases quadratically with the number of states. On the other hand, it is generally easier to attach an economic interpretation to a low-dimensional state space (e.g., a Markov chain with two states can be be used to represent bull and bear market regimes).

Rydén et al. (1998) showed that hidden Markov models can reproduce reasonably well most of the stylized facts of financial return series. However, they also argue that the model seems to be "doomed from the start" for replicating the high degree of persistence in volatility that is empirically observed. This is because, similarly to traditional stationary autoregressive moving-average models, regime-switching models based on a Markovian switching process have a short memory, that is, they can only generate an autocorrelation function that eventually decays exponentially. However, at finite lags the decay in this autocorrelation function can still potentially be quite slow. For instance, past research has shown that a time series generated with a short memory process contaminated by occasional breaks can exhibit statistical properties that are akin to those that would be obtained from a genuine long memory process (e.g., Diebold and Inoue, 2001; Granger and Hyung, 2004; Mikosch and Starica, 2004; Perron and Qu, 2010; Starica and Granger, 2005). This observation explains why several studies in financial econometrics consider models in which a low-dimensional regime-switching process is used as a way to govern time-variation in the parameters of an existing econometric model. An example of such a combination is the regime-switching generalized autoregressive conditional heteroskedasticity (GARCH) model (Gray, 1996;

Haas et al., 2004).

An alternative to these types of models is to consider regime-switching processes with a high-dimensional finite state space, such as the Markov-switching multifractal (MSM) model proposed by Calvet and Fisher (2004). These authors demonstrate that this process has the ability to generate a high degree of volatility persistence and show that it outperforms GARCH, fractionally integrated GARCH, as well as regime-switching GARCH models, when modeling exchange rate volatility. Although these empirical results offer a motivation for considering pure regime-switching specifications with a large number of states, very few models of this type have since been proposed in the literature.

Building on the MSM approach, the objective of this article is to propose a new parsimonious regime-switching volatility model with a high-dimensional finite state space: the factorial hidden Markov volatility (FHMV) model. The volatility dynamics in this model originate from the product of three components: a high-dimensional Markov chain driving volatility persistence, a jump process capable of generating non-persistent changes in volatility, and a data-driven component capturing the leverage effect. The structure of the Markov chain component shares some similarities with the structure of the MSM model, because it is constructed by multiplying a large number of independent two-state Markov chains. However, the specific formulation that we adopt leads to four important differences. First, all of our two-state Markov chains are not constrained to take identical values as in the MSM model. As a consequence, the support of the volatility distribution in the FHMV model comprises thousands of points, whereas the MSM models implemented by Calvet and Fisher (2004) only allows the volatility process to switch between at most eleven different values. Second, the transition matrix of our Markov chain component is structured in such a way that the multiplicity of the second largest eigenvalue can be greater than one. This distinctive characteristic enables us to generate a high degree of volatility persistence, which translates into a very slow decay of the autocorrelation function at finite lags. A further novelty of our approach versus the MSM model is that we allow for non-persistent jumps and integrate a leverage effect. As a final advantage, the FHMV model is specified such that only one estimation of the model is sufficient while several model estimations are required to select the optimal MSM process.

We perform an empirical analysis of fit and forecasting performance on return and realized volatility data from the Standard and Poor's 500 Index (S&P 500), the Nasdaq Composite Index (NASDAQ) and the USD/EUR exchange rate over the period 2000–2016. When modeling returns, the fit of the FHMV model is superior to the MSM model in terms of information criteria and can even surpass that of a regime-switching GARCH model with Student-t innovations. When modeling realized variances, the FHMV model dominates multiplicative error models (MEM) (Engle, 2002) and logarithmic heterogeneous autoregressive (log-HAR) processes (Corsi, 2009; Corsi and Renò, 2012) in terms of information criteria. Finally, the forecasting comparison reveals that at any horizon (up to 100 days), the root mean squared forecast errors (RMSFE) generated by the FHMV model with leverage effect are either significantly smaller or comparable in size to the smallest errors produced by the competing models.

The paper is structured as follows. Section 2 introduces the FHMV model, exposes its statistical properties and relates it to the literature. Section 3 covers model estimation. Section 4 presents the results of our empirical study. Section 5 concludes. An online supplementary appendix (SA) provides detailed proofs of the theoretical results presented in the paper as well as additional information on our modeling framework and empirical results.

2 Model definition and properties

The FHMV model is designed to fit a time series of financial returns or realized variances. Its central component is a discrete-time latent variance process denoted by $\{V_t\}$. Before defining this component in detail, we introduce the modeling framework that enables us to link it to either financial returns or realized variances.

2.1 Basic modeling framework

2.1.1 Returns

Let r_t, $t = 1, \ldots, T$, denoted by $\{r_t\}$, represent a time series of demeaned financial log-returns. As is typical in the financial econometrics literature, we model r_t as

$$r_t = \sqrt{V_t}\epsilon_t, \tag{1}$$

where $\{\epsilon_t\}$ is an independent and identically distributed (i.i.d.) innovation process with mean 0 and variance 1, which is assumed to be independent of $\{V_t\}$.

2.1.2 Realized variances

Let $\{\mathrm{RV}_t\}$ represent a time series of realized variances, computed for instance as the sum of intraday squared returns. Because the realized variance is a positive process, we choose to model it with a multiplicative error structure (Engle, 2002) as

$$\mathrm{RV}_t = V_t \eta_t, \tag{2}$$

where $\{\eta_t\}$ is a positive i.i.d. innovation process with mean 1, which is assumed to be independent of $\{V_t\}$. As argued by Engle (2002), the main advantage of the multiplicative error structure is that the variable of interest is modeled without any transformation by a process that ensures its positivity. MEM have been shown to perform well on realized volatility data by Engle and Gallo (2006), Gallo and Otranto (2015) and Lanne (2006), among others.

Remark 1. The return model considered in Equation (1) implies a MEM for squared returns as $r_t^2 = V_t \eta_t$, where, in this specific context, $\eta_t = \epsilon_t^2$.

2.2 Latent variance process

We first define the latent variance process $\{V_t\}$ without a leverage component as this allows us to study the main statistical properties of our model analytically. We model V_t as

$$V_t = \sigma^2 C_t M_t, \tag{3}$$

where $\{C_t\}$ is a Markov chain with a finite state space satisfying $\mathbb{E}\left(C_t\right) = 1$, and $\{M_t\}$ is a sequence of i.i.d. discrete random variables assumed independent of $\{C_t\}$ and that satisfies $\mathbb{E}\left(M_t\right) = 1$. As a consequence, the parameter σ^2 denotes the unconditional expectation of the latent variance process, that is, $\mathbb{E}\left(V_t\right) = \sigma^2$.

The economic interpretation that we attach to the model is one where volatility is impacted by the arrival of news in the financial market, with varying degrees of importance from day to day. The processes $\{C_t\}$ and $\{M_t\}$ are both used to represent the impacts of these news. The C_t component models news whose effect persists over time, whereas M_t captures the impact of non-persistent news and can be interpreted as a jump component. These interpretations become more apparent in Sections 2.2.1 and 2.2.2, where we define C_t and M_t, respectively.

2.2.1 Structure and interpretation of C_t

The process $\{C_t\}$ is constructed as a product of N independent two-state Markov chains, denoted by $\{C_t^{(i)}\}$, $i = 1, \ldots, N$:

$$C_t = c_0 \prod_{i=1}^{N} C_t^{(i)}, \tag{4}$$

where $c_0 = 1/\mathbb{E}\left[\prod_{i=1}^{N} C_t^{(i)}\right]$ is a normalizing constant ensuring that $\mathbb{E}\left(C_t\right) = 1$. These Markov chains are assumed to share the same 2×2 transition probability matrix (t.p.m.)

$$\boldsymbol{P} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}, \tag{5}$$

where $p \in (0,1)$. However, they do not share the same state space as we assume that $C_t^{(i)} \in \{c_i, 1\}$, where $c_1 > 1$ and

$$\begin{aligned} c_i &= (1-\theta_c) + \theta_c c_{i-1} \\ &= 1 + \theta_c^{i-1}(c_1 - 1), \quad \text{for } i = 2, \ldots, N \text{ and } \theta_c \in [0,1]. \end{aligned}$$

The normalizing constant in Equation (4) is thus obtained as $c_0 = \left[\prod_{i=1}^{N}\left(1 + \theta_c^{i-1}(c_1-1)/2\right)\right]^{-1}$. Note that $c_1 \geq c_2 \geq \ldots \geq c_N \geq 1$, which implies a hierarchical structure in the components of C_t. For instance, if we say that the component $C_t^{(i)}$ is turned ON at time t when $C_t^{(i)} = c_i$ and turned OFF when $C_t^{(i)} = 1$, then $C_t^{(1)}$ and $C_t^{(N)}$ have, respectively, the greatest and weakest impact on volatility when turned ON.

The two-state Markov chains $\{C_t^{(i)}\}$, $i = 1, \ldots, N$, are used to model the impact of news arriving in the financial market, so that when any one of these chains is turned ON, volatility increases proportionally to the news importance, measured by the value of c_i. The impact of news on volatility then persists for a number of time periods that follows a geometric distribution with parameter p; in the applications reported in Section 4, the estimated value of p is very close to 1.

Remark 2. The C_t component consists of N two-state Markov chain components and can be expressed as $\log C_t = \log c_0 + \sum_{i=1}^{N} \log C_t^{(i)}$. Because a two-state Markov chain can be represented as an AR(1) process (see for instance Hamilton, 1994, chapter 22), the logarithm of the persistent volatility component can be viewed as the sum of N autoregressive components. Interestingly, the paper by Andersen and Bollerslev (1997) proposes to model log-volatility as an aggregation of AR(1) processes and argues that (asymptotically) this structure can induce long-run dependence. Moreover, each AR(1) process is interpreted as an information arrival flow process. Consequently, the persistent volatility component in the FHMV model can be seen as a discrete version of their model. In Theorem 1 and Proposition 1, we show that it can also be effective at slowing down the decay of the autocorrelation function of $\{V_t\}$.

Remark 3. The persistent component is structured as a factorial hidden Markov model as defined in Ghahramani and Jordan (1997). In fact, factorial hidden Markov processes include multiple hidden Markov chains that evolve independently of each other and that are combined to produce the final state. Moreover, the factorial structure can be seen as a particular case of the hierarchical hidden Markov structure proposed in Fine et al. (1998), which consists in layers of hidden Markov chains. It must be emphasized that both the hierarchical and factorial models can be formulated as a standard hidden Markov model. This follows from the fact that a combination of low-dimensional Markov chains can be reproduced by a single high-dimensional Markov chain. However, hierarchical and factorial hidden Markov models remain practical representations of a hidden Markov process because they allow us to consider a large number of states more parsimoniously. A more detailed discussion on the relationship between these types of structures and the FHMV model is provided in the SA.

Following Remark 3, it can be seen that $\{C_t\}$ corresponds to a Markov chain on a state space $\mathcal{X}_C$ with 2^N elements, generated by the Kronecker product of the state spaces of $\{C_t^{(i)}\}$, $i = 1, \ldots, N$, that is, $\mathcal{X}_C = c_0 \cdot \{c_1, 1\} \otimes \{c_2, 1\} \otimes \cdots \otimes \{c_N, 1\}$. Its $2^N \times 2^N$ t.p.m., denoted by $\boldsymbol{P}_C$, is simply

$$\boldsymbol{P}_C = \boldsymbol{P}^{\otimes N},$$

where $\boldsymbol{P}^{\otimes N}$ is the Nth Kronecker power of $\boldsymbol{P}$ (the kth Kronecker power of $\boldsymbol{P}$ is defined inductively for $k \in \mathbb{N}$ by $\boldsymbol{P}^{\otimes 1} = \boldsymbol{P}$ and $\boldsymbol{P}^{\otimes k} = \boldsymbol{P} \otimes \boldsymbol{P}^{\otimes (k-1)}$, $k = 2, 3, \ldots$). Because we assume that $p \in (0, 1)$, $\boldsymbol{P}_C$ is a positive matrix (i.e., all elements of $\boldsymbol{P}_C$ are strictly positive), which implies that $\{C_t\}$ is an ergodic Markov chain with a unique stationary distribution, which we denote by $\boldsymbol{\pi}_C$. Lemma 5 in the SA implies that $\boldsymbol{\pi}_C = 2^{-N}\mathbf{1}_{2^N}$, where $\mathbf{1}_n$ is used to denote the n-dimensional column vector of ones, for $n \in \mathbb{N}$.

2.2.2 Structure and interpretation of M_t

The process $\{M_t\}$ is defined to be a sequence of i.i.d. discrete random variables with probability mass function

$$\Pr\left(M_t = m_0 \cdot m_i\right) = \begin{cases} q(N-1)^{-1}, & \text{if } i = 1, \ldots, N-1, \\ 1 - q, & \text{if } i = N, \end{cases}$$

where $q \in (0, 1)$, $m_1 > 1$,

$$\begin{aligned} m_i &= (1 - \theta_m) + \theta_m m_{i-1} \\ &= 1 + \theta_m^{i-1}(m_1 - 1), \quad \text{for } i = 2, \ldots, N-1, \end{aligned}$$

and $m_N = 1$. We assume that $\theta_m \in [0, 1]$, which implies that $m_1 \geq m_2 \geq \ldots \geq m_N = 1$, and use m_0 as a normalizing constant to ensure $\mathbb{E}\left(M_t\right) = 1$, which leads to $m_0 = \left[1 + q\frac{(m_1-1)(1-\theta_m^{N-1})}{(N-1)(1-\theta_m)}\right]^{-1}$.

We interpret $\{M_t\}$ as a process capturing the non-persistent impact on volatility of the arrival of news in the financial market. The parameter q corresponds to the probability of this type of news arriving in a given time period. This news has a multiplicative impact on volatility, given by one of the values m_1, m_2, …, m_{N-1}, chosen with equal probabilities (ON states), with m_1 representing the greatest impact and m_{N-1} the weakest impact. The probability of no news arriving is $1 - q$, which is associated with $m_N = 1$ (OFF state). In contrast to $\{C_t\}$, the impact of news generated by the $\{M_t\}$ process does not persist over time since it is an independent process. Consequently, this component of the model serves to generate non-persistent jumps of different magnitudes on volatility.

For further developments, it is convenient to express $\{M_t\}$ in the form of a Markov chain. To this end, let $\boldsymbol{\pi}_M$ be the column vector of the N component probabilities

$$\boldsymbol{\pi}_M = \left(\underbrace{\frac{q}{N-1}, \quad \dots \quad , \frac{q}{N-1}}_{(N-1) \text{ terms}}, \quad 1-q \right)'. \tag{6}$$

Then, $\{M_t\}$ can be expressed as a Markov chain with $N \times N$ t.p.m., $\boldsymbol{P}_M = \mathbf{1}_N \boldsymbol{\pi}'_M$, on the state space $\mathcal{X}_M$ with N elements, where $\mathcal{X}_M = m_0 \cdot \{m_1, m_2, \dots, m_N\}$. Because $q \in (0,1)$, $\boldsymbol{P}_M$ is a positive matrix and $\{M_t\}$ is an ergodic Markov chain with stationary distribution $\boldsymbol{\pi}_M$ (see Lemma 6 in the SA).

2.2.3 Markov chain structure of V_t

The latent variance at time t, V_t, is the product of C_t and M_t, as specified in Equation (3), hence it combines the effects on volatility of the arrival of persistent and non-persistent news in the financial market. Since $\{V_t\}$ is a product of two independent ergodic Markov chains, it is itself an ergodic Markov chain with $\left(N \cdot 2^N\right) \times \left(N \cdot 2^N\right)$ t.p.m., $\boldsymbol{P}_V = \boldsymbol{P}_C \otimes \boldsymbol{P}_M$, on the state space $\mathcal{X}_V$ with $N \cdot 2^N$ elements, where $\mathcal{X}_V = \sigma^2 \cdot \mathcal{X}_C \otimes \mathcal{X}_M$. Its stationary distribution is given by $\boldsymbol{\pi}_V = \boldsymbol{\pi}_C \otimes \boldsymbol{\pi}_M$ (see Lemma 7 in the SA). Note that although $\{V_t\}$ is potentially a high-dimensional Markov chain (e.g., for $N = 10$, the number of states is 10,240), it is parsimoniously indexed by only seven parameters, that is, $\{\sigma^2, p, q, c_1, m_1, \theta_c, \theta_m\}$.

2.2.4 Volatility persistence

It is a well known empirical fact that the volatility of returns on financial assets exhibits a high degree of persistence (e.g., Mandelbrot, 1963; Bollerslev, 1986). In the FHMV model, volatility persistence can be characterized by the speed at which $\text{Cov}(V_t, V_{t+k})$ approaches zero as k increases. Let $\boldsymbol{\upsilon}$ denote the $N \cdot 2^N$ column vector of the elements of $\mathcal{X}_V$, and let $\boldsymbol{\Upsilon}$ denote the $(N \cdot 2^N) \times (N \cdot 2^N)$ diagonal matrix with the elements of $\boldsymbol{\upsilon}$ on its diagonal (i.e., $\boldsymbol{\upsilon} = \boldsymbol{\Upsilon} \mathbf{1}_{N \cdot 2^N}$). Then, based on standard Markov chain theory (see Hamilton, 1994, chapter 22), we have

$$\text{Cov}(V_t, V_{t+k}) = \boldsymbol{\pi}'_V \boldsymbol{\Upsilon} \boldsymbol{P}_V^k \boldsymbol{\upsilon} - \left(\boldsymbol{\pi}'_V \boldsymbol{\upsilon}\right)^2 \tag{7}$$

$$= \boldsymbol{\pi}_V' \boldsymbol{\Upsilon}(\boldsymbol{P}_V^k - \mathbf{1}_{N\cdot 2^N}\boldsymbol{\pi}_V')\boldsymbol{\upsilon}, \quad k = 1, 2, \ldots,$$

and $\mathrm{Cov}(V_t, V_{t+k}) \to 0$ as $k \to \infty$.

Clearly, the rate at which the volatility tends to persist in time is directly related to the rate of convergence of the matrix $\boldsymbol{P}_V^k$ as k tends to infinity. It is well known that if γ denotes the second largest eigenvalue (in absolute value) of $\boldsymbol{P}_V$, then $|\gamma|^k$ is the dominating term in its asymptotic rate of convergence (see Poskitt and Chung, 1996). This observation led Rydén et al. (1998) to affirm that hidden Markov models "can only produce series with exponentially decaying autocorrelation functions," and that these models are therefore "doomed from the start" for replicating the high degree of persistence in volatility which is empirically observed. Although this affirmation holds asymptotically, Theorem 1 shows that the particular structure that we introduce to construct the Markov chain $\{V_t\}$, specifically the multiplication of N two-state Markov chains with identical t.p.m., offers a way to slow down the convergence of $\boldsymbol{P}_V^k$ as $k = 1, 2, \ldots$.

Theorem 1 (Rate of convergence of $\boldsymbol{P}_V$)**.** *Let $\gamma = 2p - 1$ and $\boldsymbol{\Pi}_V = \lim_{k\to\infty} \boldsymbol{P}_V^k$.*

(i) Asymptotic limit of $\boldsymbol{P}_V^k$ as $k \to \infty$:

$$\boldsymbol{\Pi}_V = \mathbf{1}_{N\cdot 2^N}\boldsymbol{\pi}_{\boldsymbol{V}}'.$$

(ii) Non-asymptotic rate of convergence of $\boldsymbol{P}_V^k$ as $k = 1, 2, \ldots$:

$$\|\boldsymbol{P}_V^k - \boldsymbol{\Pi}_V\|_\infty \leq (1 + |\gamma|^k)^N - 1, \tag{8}$$

where $\|\cdot\|_\infty$ is the maximum absolute row sum norm and, for $\gamma \in [0, 1)$,

$$\|\boldsymbol{P}_V^k - \boldsymbol{\Pi}_V\|_{\max} = \left((1 + \gamma^k)^N - 1\right) \|\boldsymbol{\pi}_V\|_\infty, \tag{9}$$

with $\|\boldsymbol{\pi}_V\|_\infty = 2^{-N} \max\{q/(N-1), 1-q\}$, where $\|\cdot\|_{\max}$ is the max norm, that is, the maximum absolute element of the given matrix.

(iii) Asymptotic rate of convergence of $\boldsymbol{P}_V^k$ as $k \to \infty$:

$$\boldsymbol{P}_V^k - \boldsymbol{\Pi}_V = O(k^{N-1}|\gamma|^k). \tag{10}$$

Remark 4. From a linear algebra standpoint, N corresponds to the algebraic multiplicity of the eigenvalue γ of the matrix $\boldsymbol{P}_V$, which is its largest eigenvalue (in absolute value) that is smaller than 1. Note that the 2×2 matrix $\boldsymbol{P}$ also has an eigenvalue of $\gamma = 2p - 1$, but its algebraic multiplicity is 1. Since N corresponds to the number of components used in the construction of $\{C_t\}$, the algebraic multiplicity of the eigenvalue γ of the matrix $\boldsymbol{P}_V$ increases by one unit each time a component is added.

Theorem 1 shows that the number of latent components N impacts the rate of convergence of $\boldsymbol{P}_V^k$ as $k = 1, 2, \ldots$. For instance, if $N = 1$, we have $\|\boldsymbol{P}_V^k - \boldsymbol{\Pi}_V\|_\infty \leq |\gamma|^k$ and $\boldsymbol{P}_V^k - \boldsymbol{\Pi}_V = O(|\gamma|^k)$. Equations (8)–(10) indicate that higher values of N generally lead to a slower decay of $\boldsymbol{P}_V^k$ as $k = 1, 2, \ldots$, and that the impact of a higher N is magnified the closer γ (or equivalently p) is to 1.

2.3 Autocovariance structure and moments

Although the Markov chain process $\{V_t\}$ exhibits a particular structure and has a high-dimensional state space, it nevertheless remains a time-homogeneous Markov chain on a finite state space. Consequently, the FHMV model presented in Sections 2.1 and 2.2 is included in the class of hidden Markov models. Accordingly, its autocovariance structure, its conditional and unconditional moments, as well as its log-likelihood function can all be computed in closed-form based on standard techniques.

2.3.1 Autocovariance structure

First, let us consider the autocovariance function of $\{r_t^2\}$ and $\{\mathrm{RV}_t\}$. Since r_t^2 and RV_t share the same multiplicative error structure (see Remark 1), the derivation of this function for these two processes is treated at once in Proposition 1 by introducing a new variable x_t that represents either r_t^2 or RV_t.

Proposition 1 (Autocovariance structure)**.** *Let $x_t = V_t \eta_t$, where V_t is defined by Equation* (3) *and $\{\eta_t\}$ is a positive i.i.d. random process with mean* 1 *and finite variance, which is assumed independent of $\{V_t\}$, and let*

$$\phi_i = \left(\frac{c_i - 1}{c_i + 1}\right)^2 = \left(\frac{\theta_c^{i-1}(c_1 - 1)}{\theta_c^{i-1}(c_1 - 1) + 2}\right)^2 \in [0, 1], \quad i = 1, \ldots, N.$$

For $k \in \mathbb{N}$, we have:

(i)

$$\operatorname{Cov}(x_t, x_{t+k}) = \operatorname{Cov}(V_t, V_{t+k}), \tag{11}$$

(ii)

$$\operatorname{Cov}(x_t, x_{t+k}) = \sigma^4 \left(\prod_{i=1}^{N} \left(1 + \phi_i \gamma^k\right) - 1 \right), \tag{12}$$

(iii)

$$\operatorname{Var}(x_t) = \sigma^4 \left(\mathbb{E}[\eta_t^2]\, m_0^2 \left(\prod_{i=1}^{N} (1 + \phi_i) \right) \left(\frac{q}{N-1} \sum_{i=1}^{N-1} m_i^2 + (1 - q) \right) - 1 \right), \tag{13}$$

(iv)

$$\operatorname{Corr}(x_t, x_{t+k}) = \frac{\prod_{i=1}^{N} \left(1 + \phi_i \gamma^k\right) - 1}{\mathbb{E}[\eta_t^2]\, m_0^2 \left(\prod_{i=1}^{N} (1 + \phi_i) \right) \left(\frac{q}{N-1} \sum_{i=1}^{N-1} m_i^2 + (1 - q) \right) - 1},$$

where $\gamma = 2p - 1$, p being the parameter of the t.p.m. defined in Equation (5)*.*

Remark 5. Equation (11) indicates that the autocovariance function of $\{r_t^2\}$ or $\{\mathrm{RV}_t\}$ decays at the same rate as that of $\{V_t\}$. Equation (7) implies that this decay is governed by the rate of convergence of the matrix $\boldsymbol{P}_V^k$ as k tends to infinity, which itself slows down when the number of components N increases (see Theorem 1). The particular structure of the latent variance process therefore offers a way to capture varying degrees of persistence in the data, and this is an important motivation for this structure. In fact, as can be seen in the empirical study of Section 4, the FHMV model very well mimics the autocorrelation structure of squared returns and realized variances.

To determine more explicitly how the number of components N impacts on the autocovariances, let us consider two FHMV models differing by only one latent component. If both models share the same parameters, σ^2, p, c_i, $i = 1, \ldots, N-1$ and $\gamma \geq 0$, then the autocovariances of the model with $N - 1$ components, denoted by $\operatorname{Cov}_{N-1}(x_t, x_{t+k})$, are always smaller than or equal to the autocovariances of the model with one extra component, denoted by $\mathrm{Cov}_N(x_t, x_{t+k})$, since we have

$$\begin{aligned}\mathrm{Cov}_N(x_t, x_{t+k}) &= \left(1+\phi_N\gamma^k\right)\sigma^4\left(\prod_{i=1}^{N-1}\left(1+\phi_i\gamma^k\right)-1\right)+\phi_N\gamma^k\sigma^4\\ &= (1+\phi_N\gamma^k)\,\mathrm{Cov}_{N-1}(x_t, x_{t+k})+\phi_N\gamma^k\sigma^4\\ &\geq \mathrm{Cov}_{N-1}(x_t, x_{t+k}).\end{aligned}$$

We remark that if the impact of the extra component on volatility is marginal, that is, $c_N \approx 1$, then $\phi_N \approx 0$ and $\mathrm{Cov}_N(x_t, x_{t+k}) \approx \mathrm{Cov}_{N-1}(x_t, x_{t+k})$. Therefore, if more components than necessary are considered in the model, these superfluous components will not artificially inflate the dependence structure.

Another interesting feature of Proposition 1 follows from Equation (13) because it shows that the excess kurtosis typically observed in financial returns can be captured either by the latent components C_t and M_t, or by $\mathbb{E}(\eta_t^2)$ (note that in the case of returns, $\mathbb{E}(\eta_t^2)$ is the fourth moment of ϵ_t).

2.3.2 Moments

Of particular interest is the conditional moment forecast of x_{t+h}, for $h = 1, 2, \ldots$, based on the available information up to time t (as in Section 2.3.1, x_t represents either r_t^2 or RV_t). To compute this forecast, one must first obtain the vector of filtered probabilities, denoted by $\boldsymbol{\xi}_{t|t}$, using standard filtering techniques developed for hidden Markov models (e.g., Hamilton, 1994, chapter 22). Let $\upsilon_1, \upsilon_2, \ldots, \upsilon_{N\cdot 2^N}$ denote the elements of $\boldsymbol{\upsilon}$, and let $\boldsymbol{\xi}_{t+h|t}$, where $h = 0, 1, \ldots$, be the $N \cdot 2^N$ column vector with elements

$$\xi_{i,t+h|t} = \Pr\left(V_{t+h} = \upsilon_i \mid \mathcal{F}_t\right), \quad i = 1, \ldots, N \cdot 2^N, \tag{14}$$

where $\mathcal{F}_t$ denotes the observed market information up to time t. These conditional forecast probabilities are directly obtained from the filtered probabilities since $\boldsymbol{\xi}'_{t+h|t} = \boldsymbol{\xi}'_{t|t}\boldsymbol{P}_V^h$ for $h = 1, 2, \ldots$. It is then simple to compute the conditional moment forecast, $\mathbb{E}\left[g(x_{t+h}) \mid \mathcal{F}_t\right]$, for any real-valued function $g(\cdot)$ from the following expression:

$$\mathbb{E}\left[g(x_{t+h}) \mid \mathcal{F}_t\right] = \sum_{i=1}^{N\cdot 2^N} \xi_{i,t+h|t}\, \mathbb{E}\left[g(\upsilon_i \eta_{t+h})\right]. \tag{15}$$

When $g(x) = x^r$, Equation (15) simplifies to $\mathbb{E}\left[x_{t+h}^r \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\eta_{t+h}^r\right] \sum_{i=1}^{N\cdot 2^N} \xi_{i,t+h|t}\, \upsilon_i^r$. Finally, to compute unconditional moments one must simply replace the probability vector $\boldsymbol{\xi}_{t+h|t}$ by the stationary distribution $\boldsymbol{\pi}_V$ (in fact, $\boldsymbol{\xi}_{t+h|t} \to \boldsymbol{\pi}_V$ as $h \to \infty$).

2.4 Relationship to the MSM model

Since the construction of the FHMV model is motivated by the success of the MSM approach of Calvet and Fisher (2004), it is instructive to relate it to the MSM model. The MSM process was initially proposed as a model for financial returns, and it thus admits the general form given in Equation (1). Its latent variance is specified as $\widetilde{V}_t = \tilde{\sigma}^2 \prod_{i=1}^N \widetilde{C}_t^{(i)}$, where for $i = 1, \ldots, N$:

$$\widetilde{C}_t^{(i)} = \begin{cases} \widetilde{C}_{t-1}^{(i)}, & \text{with probability } \tilde{p}_i, \\ \tilde{c}, & \text{with probability } (1-\tilde{p}_i)/2, \\ 2-\tilde{c}, & \text{with probability } (1-\tilde{p}_i)/2, \end{cases}$$

with $\tilde{c} \in (0,1)$, and $\tilde{p}_i = \tilde{a}^{\tilde{b}^{i-1}}$ for $i = 1, \ldots, N$, where $\tilde{a} \in (0,1)$ and $\tilde{b} \in (1,\infty)$. Note that $\mathbb{E}(\widetilde{V}_t) = \tilde{\sigma}^2$. The MSM model therefore includes four parameters, $\{\tilde{\sigma}^2, \tilde{a}, \tilde{b}, \tilde{c}\}$.

It is easily seen that $\{\widetilde{C}_t^{(i)}\}$, $i = 1, \ldots, N$, are independent two-state Markov chains defined on a common state space comprised of the values $\{\tilde{c}, 2-\tilde{c}\}$. Consequently, $\widetilde{V}_t$ can only take $N+1$ distinct values in the set $\{\tilde{\sigma}^2\, \tilde{c}^i (2-\tilde{c})^{N-i}\}_{i=0}^N$. This represents a first important difference with respect to the FHMV model. Moreover, in contrast to our approach, the Markov chains, $\{\widetilde{C}_t^{(i)}\}$, $i = 1, \ldots, N$, do not all share the same t.p.m., which implies that the structure of the MSM model does not benefit from the results of Theorem 1 and Proposition 1. In fact, Proposition 2 shows that the asymptotic rate of convergence of the MSM t.p.m., denoted by $\boldsymbol{P}_{\text{MSM}}$, is geometric and is driven by the parameter $\tilde{a}$, which also corresponds to the second largest eigenvalue of $\boldsymbol{P}_{\text{MSM}}$. Moreover, the multiplicity of this eigenvalue is equal to one.

Proposition 2 (Rate of convergence of $\boldsymbol{P}_{\mathrm{MSM}}$)**.**

(i) MSM stationary distribution : $\boldsymbol{\pi}_{MSM} = 2^{-N}\mathbf{1}_{2^N}$.

(ii) Asymptotic limit of $\boldsymbol{P}^k_{MSM}$ *as* $k \to \infty$*:* $\boldsymbol{\Pi}_{MSM} = \lim_{k\to\infty} \boldsymbol{P}^k_{MSM} = \mathbf{1}_{2^N}\boldsymbol{\pi}'_{MSM}$.

(iii) Asymptotic rate of convergence of $\boldsymbol{P}^k_{MSM}$ *as* $k \to \infty$*:* $\boldsymbol{P}^k_{MSM} - \boldsymbol{\Pi}_{MSM} = O(\tilde{a}^k)$.

Another model that is related to the MSM model is the component-driven regime-switching model of Fleming and Kirby (2013). Like the MSM model, it represents the latent variance by a product of two-state Markov chains with identical state spaces, but it allows some of these Markov chains to share the same t.p.m. However, the models considered by Calvet and Fisher (2004) and Fleming and Kirby (2013) are in practice restricted to switch between at most eleven different volatility values (for $N = 10$), while the FHMV model has the flexibility to generate a much richer support for the volatility distribution.

2.5 Relationship to an autoregressive stochastic volatility framework

Because the FHMV process can be converted into a hidden Markov model with a large number of states (see Section 2.2.3), the underlying latent variance process can be formulated as a first-order vector autoregression (see for instance Hamilton, 1994, chapter 22). More precisely, if the random vector $\mathbf{e}_t \in \mathbb{R}^{N\cdot 2^N}$ denotes a column vector with entry i equal to one if the Markov chain lies in state i at time t and zero in all other entries, then we can express the FHMV model as

$$\begin{aligned} r_t &= \sqrt{V_t}\epsilon_t, & (16)\\ V_t &= \boldsymbol{\upsilon}'\mathbf{e}_t, & (17)\\ \mathbf{e}_t &= \boldsymbol{P}'_V\mathbf{e}_{t-1} + u_t, & (18) \end{aligned}$$

where $\boldsymbol{\upsilon}$ stands for the $N\cdot 2^N$ column vector of elements of the state space of $\{V_t\}$, $\boldsymbol{P}_V$ is the transition matrix of $\{V_t\}$, and $\{u_t\}$ is a discrete martingale difference sequence. The model formulation (16)–(18) shows that the FHMV process can be represented as an autoregressive stochastic volatility model with discrete dynamics (see Cordis and Kirby, 2014, for a discrete stochastic autoregressive volatility model). While standard stochastic volatility models assume that log-volatility dynamics are driven by a Gaussian innovation, the FHMV process uses a discrete transition kernel that can potentially allow for more abrupt changes in volatility.

2.6 Leverage effect

An additional novelty of the FHMV model, that is not shared by the MSM process, is the inclusion of a time-varying leverage effect. The empirical analyses presented in Section 4 show that this component significantly enhances the in-sample fit and out-of-sample forecasting performance of the model on S&P 500 and NASDAQ data.

With a leverage effect, the latent variance specification introduced in Equation (3) is extended to include an additional component:

$$V_t = \sigma^2 C_t M_t L_t, \quad \text{where} \quad L_t = \prod_{i=1}^{N_L} L_t^{(i)},$$

and

$$L_t^{(i)} = \begin{cases} 1, & \text{if } r_{t-i} \geq 0, \\ 1 + l_i \frac{|r_{t-i}|}{\sqrt{L_{t-i}}}, & \text{if } r_{t-i} < 0, \end{cases}$$

for $i = 1, \ldots, N_L$, with $l_1 > 0$ and $l_i = \theta_l^{i-1} l_1$ for $i = 2, \ldots, N_L$, and $\theta_l \in [0, 1]$. The leverage component L_t adds two parameters, $\{l_1, \theta_l\}$, and is specified as a predictable process, that is, its value at time t is fully determined by the observed returns up to time $t - 1$. This entails that the specific value of N_L has a negligible impact on the computational burden and it can thus be chosen to be very large in applications.

Moreover, the process $\{L_t\}$ may be interpreted similarly to $\{C_t\}$ and $\{M_t\}$. For instance, if we let the returns $\{r_t\}$ represent a type of news, then we may say that the component $L_t^{(i)}$ is turned ON if $r_{t-i} < 0$ and turned OFF if $r_{t-i} \geq 0$. The impact of this component on volatility is then influenced by the importance of the news, which is a function of the magnitude of the negative return $|r_{t-i}|$ and of a multiplicative factor l_i structured to give less importance to more distant news.

3 Model estimation

Section 2.2.3 explained that the FHMV process can be recast into a hidden Markov model on the state space $\mathcal{X}_V$ with $N \cdot 2^N$ elements. Although the number of states grows quickly with the number of components N, the model can be estimated using the standard Hamilton filter, even when $N = 10$. This filter computes the filtering and predictive distributions of the state process as well as the conditional density of the observed process recursively for $t = 1, \dots, T$ as follows:

$$\begin{aligned}
\text{Observed density: } p(y_t \mid \mathcal{F}_{t-1}, \Theta) &= \sum_{V_t \in \mathcal{X}_V} p(y_t \mid V_t, \mathcal{F}_{t-1}, \Theta) p(V_t \mid \mathcal{F}_{t-1}, \Theta), \\
\text{Filtering distribution: } \quad p(V_t \mid \mathcal{F}_t, \Theta) &= \frac{p(y_t \mid V_t, \mathcal{F}_{t-1}, \Theta) p(V_t \mid \mathcal{F}_{t-1}, \Theta)}{p(y_t \mid \mathcal{F}_{t-1}, \Theta)}, \\
\text{Predictive distribution: } p(V_t \mid \mathcal{F}_{t-1}, \Theta) &= \sum_{V_{t-1} \in \mathcal{X}_V} p(V_t \mid V_{t-1}, \mathcal{F}_{t-1}, \Theta) p(V_{t-1} \mid \mathcal{F}_{t-1}, \Theta),
\end{aligned}$$

where $\{y_t\}$ denotes the observed time series (either r_t or RV_t) and Θ stands for the model parameters. The log-likelihood function is then obtained as $\log p(y_1, \dots, y_T \mid \Theta) = \sum_{t=1}^{T} \log p(y_t \mid \mathcal{F}_{t-1}, \Theta)$. To initiate the Hamilton filter, an assumption on the state distribution at time $t = 1$, $p(V_1 \mid \mathcal{F}_0, \Theta)$, must be made; in our code, it is set to the stationary distribution of the Markov chain. A MATLAB program to estimate the FHMV model is available in the supplementary material and on the corresponding author's website. In our applications, we estimated the FHMV model with $N = 10$ and the time required to carry out maximum likelihood estimation was below 30 minutes for a sample size of 4150 observations (Table 1 in the SA gives computing times required to evaluate the likelihood function as a function of N). We remark that since the predictive distribution of the jump component is constant over time, it is not necessary to track the jump states in the Hamilton filter. This implies that in practice the filter only needs to iterate over 1,024 states instead of 10,240 states when $N = 10$. Therefore, the computational burden of the FHMV model is comparable to that of the MSM process (when $N = 10$, the MSM process corresponds to a hidden Markov model over 1,024 states).

When constructing the FHMV model, we assumed that the number of components used as building blocks of $\{M_t\}$ and $\{C_t\}$ is the same and equal to N. Although nothing prevents us from considering different numbers of components in $\{M_t\}$ and $\{C_t\}$, in our view it makes sense to specify N as large as possible in both of them (up to computational constraints), because the effect of additional components on volatility, measured by the variables c_i and m_i, is structured to converge geometrically to one. Therefore, when N is large, the model has the ability to adjust itself, through the parameters c_1, m_1, θ_c and θ_m, and assign very little importance to superfluous components. Since the number of parameters does not increase with the number of components, we could also have pursued a strategy to find the optimal N. We decided not to consider such an approach because in our view, it is more practical to have only one model specification to estimate. In this respect, processes such as the MSM and GARCH(p,q) models may be considered at a disadvantage because they require a model selection procedure.

4 Applications to daily returns and realized variances

We compare the FHMV process to popular models on daily percentage log-returns and realized kernel variances (scaled by a factor of 100^2) from the S&P 500, the NASDAQ and the USD/EUR exchange rate. Daily percentage log-returns span the period extending from January 3, 2000 to June 30, 2016 (source: Federal Reserve Economic Data (FRED) database). Realized kernel variances cover the same period, except for the USD/EUR exchange rate, for which the data is available until March 3, 2009 (source: Oxford-Man Institute of Quantitative Finance).

On each data set, we estimate the FHMV model with and without leverage based on $N = 10$ (10,240 states) and $N_L = 70$. The innovation of the return process (ϵ_t) is assumed to follow a standard normal distribution, whereas the innovation of the realized variance process is assumed to follow a gamma distribution with mean 1 and shape parameter $\upsilon > 0$.

4.1 Comparison of fit

Table 1 presents estimation results for the percentage log-return data sets. The FHMV models with and without leverage (respectively, FHMV-lev and FHMV) are compared to five competitors: the MSM (Calvet and Fisher, 2004), the GARCH(1,1) (Bollerslev, 1986), the GJR-GARCH(1,1) (GJR) (Glosten et al., 1993), the two-state Markov-switching GARCH(1,1) (MS-GARCH) (Haas et al., 2004) and the two-state Markov-switching GJR-GARCH(1,1) (MS-GJR). GARCH-type

Table 1: Comparison of fit: Percentage log-returns.

	Models without leverage				Models with leverage		
Models	MSM	GARCH-t	MS-GARCH-t	FHMV	GJR-t	MS-GJR-t	FHMV-lev
Np	4	4	9	7	5	11	9
				S&P 500 ($T = 4150$)			
log-lik	−5874.1	−5870.4	−5861.3	−5862.9	−5782.0	−5770.5	**−5770.1**
AIC	−5878.1	−5874.4	−5870.3	−5869.9	−5787.0	−5781.5	**−5779.1**
BIC	−5890.8	−5887.1	−5898.8	−5892.1	**−5802.8**	−5816.3	−5807.5
				NASDAQ ($T = 4149$)			
log-lik	−7261.3	−7259.5	−7248.4	−7252.8	−7197.5	**−7175.5**	−7180.2
AIC	−7265.3	−7263.5	−7257.4	−7259.8	−7202.5	**−7186.5**	−7189.2
BIC	−7278.0	−7276.2	−7285.9	−7281.9	−7218.3	−7221.3	**−7217.7**
				USD/EUR ($T = 4147$)			
log-lik	−3762.2	−3747.5	−3740.7	−3738.6	−3745.6	−3740.5	**−3737.7**
AIC	−3766.2	−3751.5	−3749.7	**−3745.6**	−3750.6	−3751.5	−3746.7
BIC	−3778.9	**−3764.2**	−3778.2	−3767.8	−3766.4	−3786.3	−3775.1

Np: Number of parameters; log-lik: Maximum of the log-likelihood; AIC: Akaike Information criterion; BIC: Bayesian information criterion; The highest values appear in bold.

models include a Student-t innovation; this is indicated by adding "-t" to the model acronym. Model definitions are provided in the SA.

From Table 1, we observe that, in accordance with the financial econometrics literature, the inclusion of a leverage effect strongly improves the fit to stock indices, but has little impact on the exchange rate data set. Overall, the fit of the FHMV (respectively, FHMV-lev) model is comparable to that of the MS-GARCH-t (respectively, MS-GJR-t). Based on the AIC, the FHMV-lev model is preferred for the S&P 500 data set, the MS-GJR-t for the NASDAQ, and the FHMV for the USD/EUR. Based on the BIC, the FHMV-lev model is preferred for the NASDAQ. Moreover, although the MSM process was originally proposed for exchange rate series, the FHMV model substantially outperforms it in terms of information criteria.

Table 2 presents estimation results for the realized variance data sets. The competing models are: the MEM (Engle, 2002), the two-state Markov-switching MEM (MS-MEM) (Gallo and Otranto, 2015) and the log-HAR (Corsi, 2009; Corsi and Renò, 2012). These models are implemented with and without leverage; models with a leverage effect are indicated by adding "-lev" to the model acronym. Leverage in the MEM and MS-MEM is introduced as in Gallo and Otranto (2015), whereas leverage in the log-HAR is modeled as in Corsi and Renò (2012). The competing MEM models include a gamma distributed innovation with mean 1 and shape parameter $v > 0$

Table 2: Comparison of fit: Realized variances.

	Models without leverage				Models with leverage			
Models	log-HAR	MEM	MS-MEM	FHMV	log-HAR-lev	MEM-lev	MS-MEM-lev	FHMV-lev
Np	5	4	9	8	8	5	11	10
					S&P 500 ($T = 4120$)			
log-lik	−1196.4	−1465.0	−1234.0	−1148.4	−1033.9	−1366.4	−1120.7	**−956.9**
AIC	−1201.4	−1469.0	−1243.0	−1156.4	−1041.9	−1371.4	−1131.7	**−966.9**
BIC	−1217.2	−1481.7	−1271.5	−1181.7	−1067.2	−1387.2	−1166.5	**−998.5**
					NASDAQ ($T = 4124$)			
log-lik	−1444.9	−1669.0	−1464.6	−1427.2	−1256.7	−1596.7	−1389.8	**−1225.5**
AIC	−1449.9	−1673.0	−1473.6	−1435.2	−1264.7	−1601.7	−1400.8	**−1235.5**
BIC	−1465.7	−1685.6	−1502.1	−1460.5	−1290.0	−1617.5	−1435.6	**−1267.2**
					USD/EUR ($T = 2328$)			
log-lik	1293.5	1150.0	1266.7	1325.3	1307.7	1151.8	1269.5	**1335.5**
AIC	1288.5	1146.0	1257.7	1317.3	1299.7	1146.8	1258.5	**1325.5**
BIC	1274.1	1134.5	1231.8	1294.3	1276.7	1132.4	1226.8	**1296.7**

Np: Number of parameters; log-lik: Maximum of the log-likelihood; AIC: Akaike Information criterion; BIC: Bayesian information criterion; The highest values appear in bold.

whereas the log-HAR processes use a normal innovation. Model definitions are provided in the SA. Overall, we observe that estimation results strongly favor the FHMV-lev model for all data sets.

4.2 Value-added of the jump and leverage components

Table 3 shows how the log-likelihood (evaluated at the MLE) and the BIC of the FHMV model increase when the jump component and the leverage effect are added. Overall, these two components improve the log-likelihood by a greater margin when the model is fitted to realized variances than to returns. This observation therefore partly explains why the model shows a greater outperformance for the realized variance data sets in the previous section.

As expected, the contribution of the leverage component is very strong for S&P 500 and NASDAQ data, and insignificant for the USD/EUR exchange rate according to the BIC. Moreover, we note that the contribution of the jump component is always significant when evaluated with respect to the BIC. We believe that this component turns out to be more important for the realized variance series because the conditional variance dynamics are more directly observed in that case, and abrupt changes are therefore easier to detect. In contrast, squared log-returns are a relatively

Table 3: Contribution of the jump and leverage components in the FHMV model.

Percentage log-returns			
	S&P 500	NASDAQ	USD/EUR
FHMV w/o jump	−5890.6	−7279.2	−3762.5
Increase in log-likelihood with respect to FHMV w/o jump			
FHMV	27.7	26.5	23.85
FHMV-lev w/o jump	92.8	76.9	0.2
FHMV-lev	120.6	99.0	24.8
Increase in BIC with respect to FHMV w/o jump			
FHMV	19.4	18.0	15.5
FHMV-lev w/o jump	84.5	68.6	−8.1
FHMV-lev	103.9	82.3	8.1
Realized variances			
	S&P 500	NASDAQ	USD/EUR
FHMV w/o jump	−1209.8	−1459.8	1274.1
Increase in log-likelihood with respect to FHMV w/o jump			
FHMV	61.4	32.6	51.2
FHMV-lev w/o jump	146.0	182.9	5.3
FHMV-lev	252.9	234.3	61.4
Increase in BIC with respect to FHMV w/o jump			
FHMV	53.1	24.4	42.9
FHMV-lev w/o jump	137.7	174.6	−3.0
FHMV-lev	236.3	217.6	44.8

noisy proxy of conditional variance and this fact renders the identification of sharp changes in volatility more difficult.

4.3 Analysis of the fit to S&P 500 data

4.3.1 Estimated parameters

Table 4 reports parameter estimates for the FHMV-lev model fitted to S&P 500 returns and realized variances. For interpreting the values, remember that when a component $C_t^{(i)}$ in the model is turned ON, it has a multiplicative impact of c_i on the variance V_t. The jump component on the other hand has an overall multiplicative effect of $m_i m_0$.

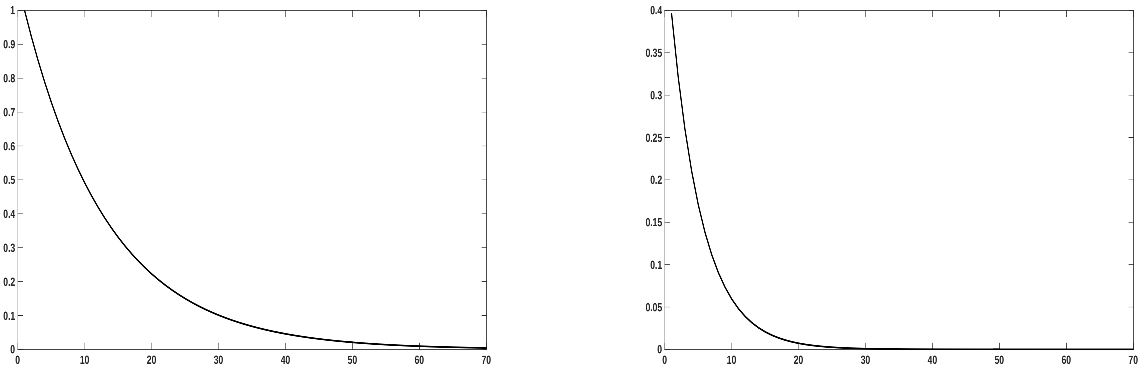
With respect to the model for returns, we observe that each component $C_t^{(i)}$ persists for an average of two years (i.e., $1/(1-p)$ days) when turned ON, and that the strongest component can double the variance value. Moreover, jumps that increase the variance are approximately as

Table 4: S&P 500: Maximum likelihood estimates of the FHMV-lev model.

Percentage log-returns									
Constant component : $\sigma^2 = 0.22$									
Markov chain component : $\theta_c = 0.51, c_1 = 1.99, p = 0.9986$									
c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
1.99	1.50	1.26	1.13	1.07	1.03	1.02	1.01	1.00	1.00
Jump component : $\theta_m = 0.87, m_1 = 23.55, q = 0.93$									
m_1m_0	m_2m_0	m_3m_0	m_4m_0	m_5m_0	m_6m_0	m_7m_0	m_8m_0	m_9m_0	$m_{10}m_0$
1.69	1.48	1.31	1.15	1.02	0.90	0.79	0.70	0.62	0.07
Leverage component : $\theta_l = 0.92, l_1 = 1.00$									
Realized variances									
Constant component : $\sigma^2 = 0.51$									
Markov chain component : $\theta_c = 0.81, c_1 = 2.19, p = 0.9897$									
c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
2.19	1.97	1.78	1.64	1.52	1.42	1.34	1.28	1.22	1.18
Jump component : $\theta_m = 0.75, m_1 = 3.45, q = 0.12$									
m_1m_0	m_2m_0	m_3m_0	m_4m_0	m_5m_0	m_6m_0	m_7m_0	m_8m_0	m_9m_0	$m_{10}m_0$
3.08	2.53	2.12	1.81	1.58	1.40	1.27	1.18	1.11	0.89
Leverage component : $\theta_l = 0.81, l_1 = 0.40$									

frequent as those that decrease it (i.e., $\Pr(M_t > 1) = 0.52$). When looking at the model for realized variances, the impact of persistent news lasts on average for 100 days and jumps that increase the variance are relatively less frequent (i.e., $\Pr(M_t > 1) = 0.12$).

Figure 1 illustrates the leverage coefficients l_i for $i = 1, \ldots, 70$. We observe that until around 60 (respectively, 20), past negative returns are relevant to build the leverage component in the model for returns (respectively, realized variances). We can interpret this long-lasting impact as the time needed for the financial market to completely react to a negative return.



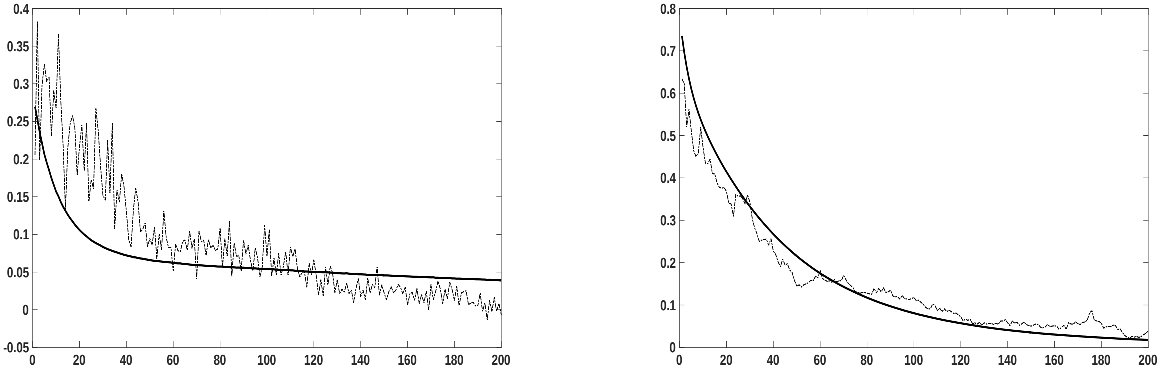
(a) Percentage log-returns

(b) Realized variances

Figure 1: S&P 500 : Leverage coefficients l_i for $i = 1, \ldots, 70$ in the FHMV-lev model.

4.3.2 Autocorrelation structure

Figure 2 plots the empirical autocorrelations of the squared percentage log-returns and of the realized variances against the theoretical ones implied by the estimated FHMV-lev model (the autocorrelations of the FHMV-lev model were computed by simulation). We observe a long-lasting volatility persistence that is reasonably well tracked by the model, especially for realized variances.



(a) Squared percentage log-returns

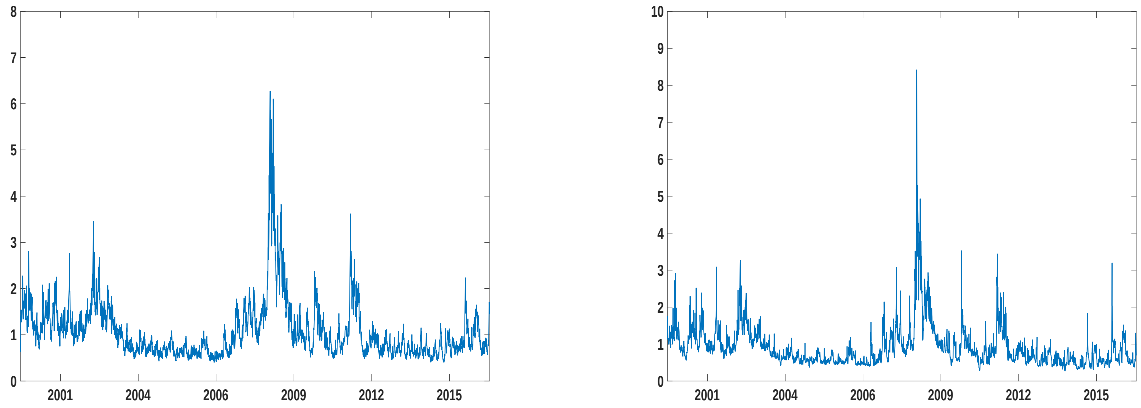
(b) Realized variances

Figure 2: S&P 500 : Theoretical autocorrelations implied by the FHMV-lev model (solid line) against empirical autocorrelations (dashed line).

4.3.3 Inference on V_t

The fact that the Markov chain and jump components in the FHMV model imply a discrete process for the latent variance may raise some concerns about the flexibility of volatility dynamics in the model. Figure 3 illustrates the median of the distribution of the inferred smoothed conditional volatility at each time point (i.e., the median of $p(\sqrt{V_t} \mid \mathcal{F}_T)$) in the estimated FHMV-lev models. We observe that volatility evolves as if it was generated by a stochastic volatility model with a continuous state space.

Figure 4 shows the time periods when the first three $C_t^{(i)}$ components in the FHMV-lev model are likely to be turned ON (i.e., when $\Pr(C_t^{(i)} = c_i \mid \mathcal{F}_T) \geq 0.5$). With respect to the model for returns, we observe that the component having the strongest impact on volatility is likely to be active only during the major stock market crashes of 2000–2016 (dot-com crisis in 2000, subprime mortgage crisis in 2008 and European sovereign debt crisis in 2011). For the realized volatility

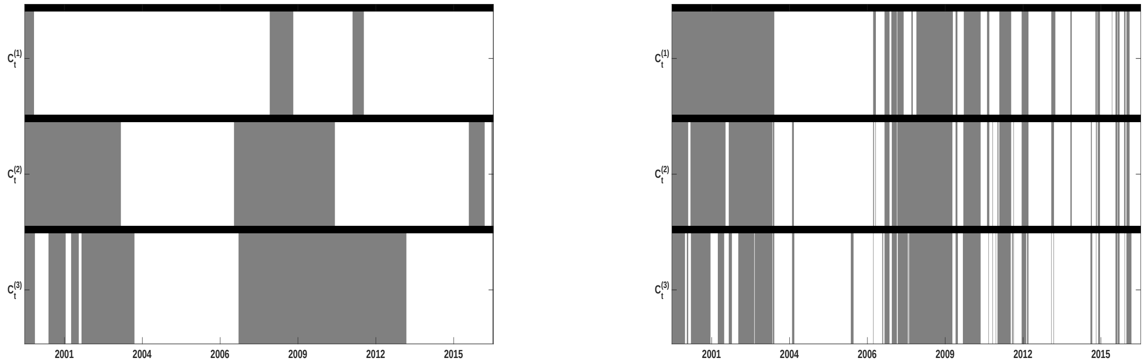


(a) Percentage log-returns

(b) Realized variances

Figure 3: S&P 500: Inferred smoothed conditional volatilities in the FHMV-lev model.

model, the first three components are often turned ON simultaneously during periods of market stress, which suggests that a dependence structure between the components could be incorporated into the model.



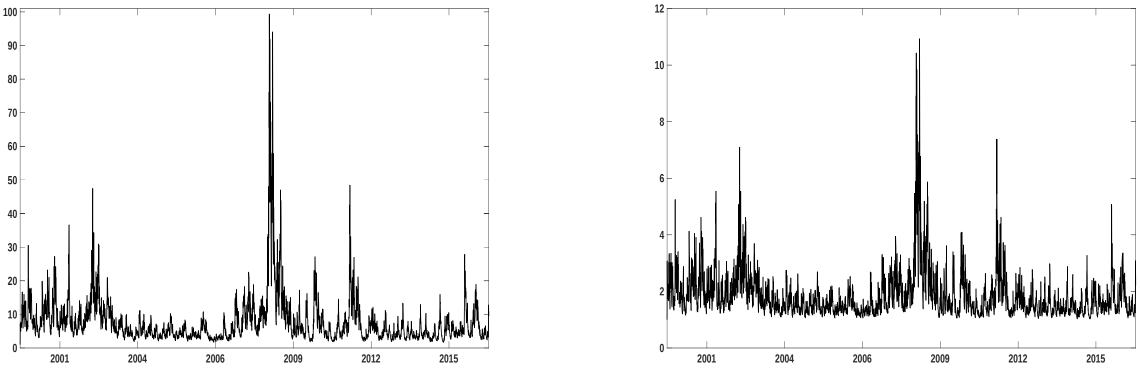
(a) Percentage log-returns

(b) Realized variances

Figure 4: S&P 500 : Time periods when the first three $C_t^{(i)}$ components in the FHMV-lev model are likely to be turned ON (i.e., when $\Pr(C_t^{(i)} = c_i \mid \mathcal{F}_T) \geq 0.5$). The component with the largest impact is displayed on the top. The second and third components are shown in the middle and at the bottom, respectively.

4.3.4 Analysis of the leverage effect

Figure 5 shows the values taken by the leverage effect component L_t over time. We observe that its effect is very strong during the subprime mortgage crisis. Since this component is specified in a non-traditional way, it can be questioned whether it really corresponds to the so-called leverage



(a) Percentage log-returns

(b) Realized variances

Figure 5: S&P 500 : Leverage effect component (L_t) over time in the FHMV-lev model.

effect captured by standard volatility models, such as the GJR model. To investigate this issue, note first that, assuming standardized innovations, the GJR conditional variance process can be decomposed as follows:

$$\begin{aligned}\sigma_t^2 &= \omega + \left(\alpha + \delta \mathbb{1}_{\{r_{t-1}<0\}}\right) r_{t-1}^2 + \beta \sigma_{t-1}^2, \\ &= \omega \sum_{i=1}^{t} \beta^{i-1} + \alpha \sum_{i=1}^{t} \beta^{i-1} r_{t-i}^2 + \underbrace{\delta \sum_{i=1}^{t} \beta^{i-1} \mathbb{1}_{\{r_{t-i}<0\}} r_{t-i}^2}_{L_t^{GJR}} + \beta^t \sigma_0^2. \end{aligned} \tag{19}$$

The decomposition (19) isolates the contribution of the leverage effect to the variance dynamics in the GJR model; its impact at time t corresponds to $L_t^{GJR} = \delta \sum_{i=1}^{t} \beta^{i-1} \mathbb{1}_{\{r_{t-i}<0\}} r_{t-i}^2$. Therefore, the L_t^{GJR} component depends on several previous squared returns and the influence of the ith lagged squared return is scaled by the coefficient $l_i^{GJR} = \delta \beta^{i-1}$, which geometrically decays over time. Although there is a clear connection between L_t and L_t^{GJR}, as well as between l_i and l_i^{GJR}, the impacts of these components on volatility cannot be directly compared. This is due to the fact that the leverage component in the FHMV model is specified as a multiplicative component, whereas the impact of the GJR leverage component on variance is additive. Nevertheless, the correlation coefficient between L_t and L_t^{GJR} is equal to 0.92, which confirms that both components capture a similar effect.

4.4 Forecasting performance

We carry out a forecasting exercise over the last three years (756 trading days) of the data sample periods in order to compare the predictions of the FHMV models and of some competitors on short- and long-run forecasting periods. Each time we move forward by one day in the in-sample period, the models are re-estimated, and cumulative variance forecasts, $\sum_{i=1}^{h} \hat{V}_{t+i}$, where $\hat{V}_{t+i} = \mathbb{E}[x_{t+i} \mid \mathcal{F}_t]$, are computed for time horizons of $h = 1, 5, 10, 25, 50, 75$ and 100. These forecasts are then compared to their associated observed values $\sum_{i=1}^{h} x_{t+i}$, where x_t denotes either the realized variance or the squared percentage log-return. The comparison of forecasts is based on the (normalized) root mean squared forecast error (RMSFE) defined as

$$\text{RMSFE for horizon } h = \sqrt{\frac{1}{756-h+1} \sum_{t=0}^{756-h} \left(\frac{1}{h} \sum_{i=1}^{h} \hat{V}_{t+i} - \frac{1}{h} \sum_{i=1}^{h} x_{t+i} \right)^2}, \qquad (20)$$

where $t = 0$ represents the end of the in-sample period.

Note that return predictions are needed to produce multi-step realized variance forecasts for models that include a leverage effect. To this end, we assume that future percentage log-returns are i.i.d. $N(\hat{\mu}_t, \hat{\sigma}_t^2)$, where $\hat{\mu}_t$ and $\hat{\sigma}_t^2$ denote respectively the empirical mean and variance over the last three years of the in-sample period. The forecasting of the leverage effect could be improved by considering a bivariate model for returns and realized variances, an extension we leave to further research.

Tables 5 and 6 show the forecasting performance of all models on percentage log-return and realized variance data sets, respectively. The statistical significance of differences in forecasting errors is assessed based on the Diebold-Mariano (DM) test (Diebold and Mariano, 2002). For this purpose, the GARCH-t and GJR-t models act as benchmarks for returns and the log-HAR and log-HAR-lev models for realized variances.

For the S&P 500 and NASDAQ return series, the FHMV-lev model produces the smallest RMSFE (with two exceptions; the GJR-t and MS-GJR-t models perform slightly better for the NASDAQ data set at forecast horizons 1 and 5). The differences between the RMSFE of the FHMV-lev model and those of its competitors increase noticeably with the forecast horizon. For instance, its forecasting performance is found to be superior with respect to the GJR-t model at a 5% or 10% level at horizons larger than 10 days (for the NASDAQ and S&P 500). For the

Table 5: Percentage log-returns: RMSFE computed over the last three years of the data sample period.

	Horizon (h)	1	5	10	25	50	75	100
					S&P 500			
Without leverage	GARCH-t	1.38	0.89	0.77	0.68	0.65	0.62	0.57
	MSM	1.38	0.89	0.77	0.68	0.64	0.61	0.56
	MS-GARCH-t	1.38	0.88	0.77	0.68	0.64	0.60	0.55
	FHMV	1.39	0.90	0.79	0.72	0.70	0.69	0.64
With leverage	GJR-t	1.33	0.85	0.77	0.71	0.67	0.63	0.56
	MS-GJR-t	1.35	0.84	0.73	0.59*	0.50**	0.45	0.40**
	FHMV-lev	**1.33**	**0.81**	**0.69**	**0.56***	**0.48****	**0.42****	**0.37****
					NASDAQ			
Without leverage	GARCH-t	2.05	1.23	1.08	0.95	0.91	0.87	0.76
	MSM	2.04	1.23	1.08	0.94	0.91	0.90	0.85
	MS-GARCH-t	2.06	1.28	1.16	1.12	1.20	1.26	1.25
	FHMV	2.06	1.26	1.11	1.01	1.00	1.00	0.91
With leverage	GJR-t	1.99	1.19	1.06	1.01	1.02	1.01	0.92
	MS-GJR-t	**1.96**	**1.13**	0.99*	0.87	0.77*	0.70**	0.64*
	FHMV-lev	2.00	1.14	**0.96**	**0.77****	**0.65****	**0.56****	**0.47****
					USD/EUR			
Without leverage	GARCH-t	0.66	0.30	0.22	0.16	0.14	0.14	0.16
	MSM	0.66	0.30	0.22	0.16	0.14	0.14	0.15
	MS-GARCH-t	0.66**	0.30**	0.21**	0.15**	0.12**	0.12**	0.13**
	FHMV	0.66**	0.30**	0.21**	0.15*	0.13**	0.13	0.14
With leverage	GJR-t	0.66	0.30	0.21	0.15	0.13	0.14	0.15
	MS-GJR-t	**0.66****	**0.29***	**0.21***	**0.14****	**0.11****	**0.11****	**0.12****
	FHMV-lev	0.66	0.30	0.21	0.15	0.13	0.14	0.15

A star means that the squared forecasting error is significantly smaller than that of the benchmark process (GARCH-t for models without leverage, GJR-t for models with leverage) at the 10% level when using the DM test. A double star stands for a 5% significance level. The smallest RMSFE appear in bold.

USD/EUR return data set, we note that the FHMV model significantly outperforms the GARCH-t model at horizons smaller or equal to 50 days at a 5 or 10% level.

With respect to realized variances (Table 6), the FHMV-lev model produces the smallest RMSFE at all horizons for S&P 500 data. The magnitude of the outperformance increases with the forecast horizon with respect to the MEM-lev and MS-MEM-lev models, but the log-HAR-lev model also performs well and the differences between this process and the FHMV-lev model are not found to be significantly different at 5 or 10% levels. For NASDAQ realized variances, the FHMV-lev process generates the smallest RMSFE for horizons up to 25 days and its predictions at 1 and 5 days are found to be significantly better than those of the log-HAR-lev model. For USD/EUR realized variances, all models without leverage perform similarly and no significant differences were detected.

Table 6: Realized variances: RMSFE computed over the last three years of the data sample period.

	Horizon (h)	1	5	10	25	50	75	100
					S&P 500			
Without leverage	log-HAR	0.78	0.54	0.48	0.43	0.42	0.40	0.36
	MEM	0.82	0.64	0.62	0.61	0.63	0.66	0.67
	MS-MEM	0.82	0.64	0.62	0.62	0.66	0.70	0.72
	FHMV	0.78	0.56	0.52	0.50	0.53	0.57	0.59
With leverage	log-HAR-lev	0.80	0.54	0.47	0.39	0.34	0.29	0.24
	MEM-lev	0.83	0.66	0.63	0.59	0.57	0.54	0.52
	MS-MEM-lev	0.83	0.66	0.64	0.60	0.58	0.57	0.55
	FHMV-lev	**0.75**	**0.52**	**0.46**	**0.39**	**0.33**	**0.28**	**0.24**
					NASDAQ			
Without leverage	log-HAR	0.58	0.44	0.41	0.40	0.42	0.42	0.38
	MEM	0.61	0.53	0.53	0.56	0.64	0.72	0.78
	MS-MEM	0.62	0.53	0.54	0.56	0.65	0.73	0.79
	FHMV	0.58	0.46	0.45	0.51	0.64	0.75	0.83
With leverage	log-HAR-lev	0.59	0.44	0.40	0.35	**0.31**	**0.26**	**0.21**
	MEM-lev	0.58	0.48	0.48	0.48	0.53	0.57	0.60
	MS-MEM-lev	0.59	0.49	0.49	0.51	0.57	0.63	0.67
	FHMV-lev	**0.56****	**0.42***	**0.39**	**0.34**	0.31	0.27	0.23
					USD/EUR			
Without leverage	log-HAR	**0.30**	**0.27**	**0.25**	**0.26**	0.31	0.35	0.38
	MEM	0.30	0.27	0.25	0.27	**0.30**	0.34	0.37
	MS-MEM	0.30	0.27	0.25	0.27	0.32	0.36	0.38
	FHMV	0.31	0.27	0.25	0.28	0.33	0.37	0.39
With leverage	log-HAR-lev	0.31	0.27	0.25	0.27	0.33	0.37	0.39
	MEM-lev	0.31	0.28	0.25	0.27	0.30	**0.34**	**0.37**
	MS-MEM-lev	0.31	0.28	0.25	0.27	0.32	0.36	0.38
	FHMV-lev	0.33	0.29	0.29	0.33	0.38	0.40	0.40

A star means that the squared forecasting error is significantly smaller than that of the benchmark process (log-HAR for models without leverage and log-HAR-lev for models with leverage) at the 10% level when using the DM test. A double star stands for a 5% significance level. The smallest RMSFE appear in bold.

5 Conclusion

We propose the FHMV model, a new volatility process that is suited for financial returns or realized variances. We specify the latent variance process as a high-dimensional Markov chain constructed from the product of three components that can be economically interpreted. In particular, the jump process captures the reaction of the financial market to non-persistent news whereas the Markov chain component models the arrival of news with a long-lasting impact. The last component generates a leverage effect and its specification differs from what is typically found in the literature. These three processes are parsimoniously specified and, altogether, create a continuum of volatility states. We derive the moments of the process and show that the autocovariance function can exhibit a slower decay than in traditional hidden Markov models thanks to the multiplicity of the second largest eigenvalue of the t.p.m. being greater than one. This property seems beneficial empirically as we show that the FHMV model dominates the MSM process in terms of information criteria (AIC and BIC) on return data from the USD/EUR exchange rate. It also compares favorably with the MS-GJR-t model on S&P 500 and NASDAQ return data sets. Moreover, on the corresponding realized variance series, the fit of the FHMV process particularly stands out versus popular realized variance models (i.e., log-HAR, log-HAR-lev, MEM, MEM-lev, MS-MEM and MS-MEM-lev). Regarding volatility forecasting performance, the FHMV process competes well with its competitors on short-run horizons (less than 25 days). At middle to long-run horizons, it significantly improves over the GJR-t model on S&P 500 and NASDAQ return data sets. With respect to realized variances, the FHMV process generally outperforms MEM and compares similarly with the log-HAR model.

We view this volatility modeling attempt with a high-dimensional hidden Markov chain as very promising since many extensions can be entertained. We could for instance add a fourth component to take into account the trading volume or we could introduce correlated components since the diverse news seem to be related. Additionally, a multivariate extension in the spirit of Calvet et al. (2006) could be undertaken.

6 Supplementary materials

The supplementary materials include a MATLAB program for estimating the FHMV model and an online appendix. Section 1 of this appendix provides a discussion of hierarchical and factorial hidden Markov models in the context of volatility modeling, with some economic interpretations. Section 2 contains the proofs of Theorem 1 and Propositions 1 and 2 of the paper. Section 3 discusses some computational aspects associated with the estimation of the FHMV model. Sections 4 and 5 describe, respectively, the competing return and realized variance models used in the empirical study.

References

Andersen, T. G. and Bollerslev, T. (1997). Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns. *The Journal of Finance*, 52(3):975–1005.

Ang, A. and Bekaert, G. (2002). Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182.

Bauwens, L., Dufays, A., and Rombouts, J. V. K. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*, 178(part 3):508–522.

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

Calvet, L. E. and Fisher, A. J. (2004). How to forecast long-run volatility: Regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, 2(1):49–83.

Calvet, L. E., Fisher, A. J., and Thompson, S. B. (2006). Volatility comovement: a multifrequency approach. *Journal of Econometrics*, 131(1-2):179–215.

Cordis, A. S. and Kirby, C. (2014). Discrete stochastic autoregressive volatility. *Journal of Banking & Finance*, 43:160–178.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Corsi, F. and Renò, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modeling. *Journal of Business & Economic Statistics*, 30(3):368–380.

Dai, Q., Singleton, K. J., and Yang, W. (2007). Regime shifts in a dynamic term structure model of U.S. Treasury bond yields. *The Review of Financial Studies*, 20(5):1669–1706.

Diebold, F. X. and Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, 105(1):131–159.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Engle, R. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, 17(5):425–446.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1-2):3–27.

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 32(1):41–62.

Fleming, J. and Kirby, C. (2013). Component-driven regime-switching volatility. *Journal of Financial Econometrics*, 11(2):263–301.

Gallo, G. M. and Otranto, E. (2015). Forecasting realized volatility with changing average levels. *International Journal of Forecasting*, 31(3):620–634.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, 29(2):245–273.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801.

Goldfeld, S. M. and Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1):3–16.

Granger, C. W. and Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance*, 11(3):399–421.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62.

Haas, M., Mittnik, S., and Paolella, M. S. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, 2(4):493–530.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press, New Jersey.

Lanne, M. (2006). A mixture multiplicative error model for realized volatility. *Journal of Financial Econometrics*, 4(4):594–616.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, 36:394–419.

Mikosch, T. and Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics*, 86(1):378–390.

Perron, P. and Qu, Z. (2010). Long-memory and level shifts in the volatility of stock market return indices. *Journal of Business & Economic Statistics*, 28(2):275–290.

Poskitt, D. S. and Chung, S.-H. (1996). Markov chain models, time series analysis and extreme value theory. *Advances in Applied Probability*, 28(2):405–425.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, 13:217–244.

Starica, C. and Granger, C. (2005). Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87:503–522.