Forecasting risk with Markov–switching GARCH models: A large–scale performance study[☆]

David Ardia[a,b,*], Keven Bluteau[a,c], Kris Boudt[c,d], Leopoldo Catania[e]

[a] *Institute of Financial Analysis, University of Neuchâtel, Neuchâtel, Switzerland*
[b] *Department of Finance, Insurance and Real Estate, Laval University, Québec City, Canada*
[c] *Solvay Business School, Vrije Universiteit Brussel, Belgium*
[d] *School of Business and Economics, Vrije Universiteit Amsterdam, The Netherlands*
[e] *Department of Economics and Business Economics, Aarhus University and CREATES, Denmark*

Abstract

We perform a large–scale empirical study to compare the forecasting performance of single–regime and Markov–switching GARCH (MSGARCH) models from a risk management perspective. We find that, for daily, weekly, and ten–day equity log–returns, MSGARCH models yield more accurate Value–at–Risk, Expected Shortfall, and left–tail distribution forecasts than their single–regime counterpart. Also, our results indicate that accounting for parameter uncertainty improves left–tail predictions, independently of the inclusion of the Markov–switching mechanism.

Keywords: GARCH, MSGARCH, forecasting performance, large–scale study, Value–at–Risk, Expected Shortfall, risk management

[☆] An earlier version of this paper was circulated under the title "Forecasting performance of Markov–switching GARCH models: A large–scale empirical study". We are grateful to the Editor (Esther Ruiz), the Associate Editor, and two anonymous referees for their useful comments which improved the paper significantly. We thank Samuel Borms, Peter Carl, Dirk Eddelbuettel, Richard Gerlach, Lennart Hoogerheide, Eliane Maalouf, Brian Peterson, Enrico Schumann, Denis–Alexandre Trottier, and participants at the Quant Insights 2017 (London), the R/Finance 2017 (Chicago), the 37th International Symposium on Forecasting (Cairns), and UseR 2017 (Brussels). We acknowledge Industrielle–Alliance, International Institute of Forecasters, Google Summer of Code 2016 and 2017, FQRSC (Grant # 2015-NP-179931) and Fonds de Donations at the University of Neuchâtel for their financial support. We thank Félix–Antoine Fortin and Calcul Québec (clusters Briaree, Colosse, Mammouth and Parallèle II) as well as Laurent Fastnacht and the Institute of Hydrology at the University of Neuchâtel (cluster Galileo) for computational support. All computations have been performed with the R package MSGARCH (Ardia et al., 2017a,b) available from the CRAN repository at `https://cran.r-project.org/package=MSGARCH`.

[*] Corresponding author. University of Neuchâtel, Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland. Phone: +41 32 718 1365.

Email addresses: `david.ardia@unine.ch` (David Ardia), `keven.bluteau@unine.ch` (Keven Bluteau), `kris.boudt@vub.be` (Kris Boudt), `leopoldo.catania@econ.au.dk` (Leopoldo Catania)

1. Introduction

Under the regulation of the Basel Accords, risk managers of financial institutions need to rely on state–of–the–art methodologies for monitoring financial risks (Board of Governors of the Federal Reserve Systems, 2012). Clearly, the use of a regime–switching time–varying volatility model and Bayesian estimation methods can be considered to be state–of–the–art, but many academics and practitioners also consider the single–regime volatility model and the use of frequentist estimation via Maximum Likelihood (ML) as state–of–the–art. Risk managers disagree whether the computational complexity of a regime–switching model and the Bayesian estimation method pay off in terms of a higher accuracy of their financial risk monitoring system. We study this question for monitoring the individual risks of a large number of financial assets.

Among the various building–blocks of any risk management system, the specification of the conditional volatility process is key, especially for short–term horizons (McNeil et al., 2015). Research on modeling volatility using time series models has proliferated since the creation of the original ARCH model by Engle (1982) and its generalization by Bollerslev (1986). From there, multiple extensions of the GARCH scedastic function have been proposed to capture additional stylized facts observed in financial markets, such as nonlinearities, asymmetries, and long–memory properties; see Engle (2004) for a review. These so–called GARCH–type models are today essential tools for risk managers.

An appropriate risk model should be able to accommodate the properties of financial returns. Recent academic studies show that many financial assets exhibit structural breaks in their volatility dynamics and that ignoring this feature can have large effects on the precision of the volatility forecast (see, *e.g.*, Lamoureux and Lastrapes, 1990; Bauwens et al., 2014). As noted by Danielsson (2011), this shortcoming in the individual forecasting systems can have systemic consequences. He refers to these single–regime volatility models as one of the culprits of the great financial crisis: *"(...) the stochastic process governing market prices is very different during times of stress compared to normal times. We need different models during crisis and non–crisis and need to be careful in drawing conclusions from non–crisis data about what happens in crises and vice versa".*

A way to address the *switch* of model's behavior is provided by Markov–switching GARCH models (MSGARCH) whose parameters can change over time according to a discrete latent (*i.e.*, unobservable) variable. These models can quickly adapt to variations in the unconditional volatility level, which improves risk predictions (see, *e.g.*, Marcucci, 2005; Ardia, 2008).

Initial studies on Markov–switching autoregressive heteroscedastic models applied to financial times series focus on ARCH specifications and thus omit a lagged value of the conditional variance in the variance equation (Cai, 1994; Hamilton and Susmel, 1994). The use of ARCH instead of GARCH dynamics leads to computational tractability in the likelihood calculation. Indeed, Gray (1996) shows that, given a Markov chain with K regimes and T observations, the evaluation of the likelihood of a Markov–switching model with general GARCH dynamics requires the integration over all K^T possible paths, rendering the estimation infeasible. While this difficulty is not present in ARCH specifications, the use of lower order GARCH models tends to offer a more parsimonious representation than higher order ARCH models. Gray (1996), Dueker (1997) and Klaassen (2002) tackle the *path dependence problem* of MSGARCH through approximation, by collapsing the past regime–specific conditional variances according to ad–hoc schemes.[1] An alternative approach is provided by Haas et al. (2004), who let the GARCH processes of each state evolve *in parallel* and thus independently of the GARCH process in the other states. Besides avoiding the path dependence problem, their model allows for a clear–cut interpretation of the variance dynamics in each regime. In our study, we consider the model by Haas et al. (2004) for these reasons.

The first contribution of our paper is to test if, indeed, MSGARCH models provide risk managers with useful tools that can improve their volatility forecasts.[2] To answer this question, we perform a large–scale empirical analysis in which we compare the risk forecasting performance of single–

[1] Also more recent studies address this problem; for instance, Augustyniak (2014) relies on a Monte Carlo EM algorithm with importance sampling.

[2] Our study focuses exclusively on GARCH and MSGARCH models. GARCH is the workhorse model in financial econometrics and has been investigated for decades. It is widely used by practitioners and academics; see for instance Bams et al. (2017) and Herwartz (2017). MSGARCH is the most natural and straightforward extension to GARCH. Alternative conditional volatility models include stochastic volatility models (Taylor, 1994; Jacquier et al., 1994), realized measure–based conditional volatility models such as HEAVY (Shephard and Sheppard, 2010) or Realized GARCH (Hansen et al., 2011), or even combinations of these (Opschoor et al., 2017). Note finally that our study only considers the (1,1)–lag specification for the GARCH and MSGARCH models. While there is a clear computational cost of considering higher orders for (MS)GARCH model specifications, the payoff in terms of improvement in forecasting precision may be low. In fact, several studies have shown that increasing the orders does not lead to a substantial improvement of the forecasting performance in case of predicting the conditional variance of asset returns (see, *e.g.*, Hansen and Lunde, 2005). We leave all these investigations for further research.

regime and Markov–switching GARCH models. We take the perspective of a risk manager working for a fund manager and conduct our study on the daily, weekly and ten–day log–returns of a large universe of stocks, equity indices, and foreign exchange rates. Thus, in contrast to Hansen and Lunde (2005), who compare a large number of GARCH–type models on a few series, we focus on a few GARCH and MSGARCH models and a large number of series. For single–regime and Markov–switching specifications, the scedastic specifications we consider account for different reactions of the conditional volatility to past asset returns. More precisely, we consider the symmetric GARCH model (Bollerslev, 1986) as well as the asymmetric GJR model (Glosten et al., 1993). These scedastic specifications are integrated into the MSGARCH framework with the approach of Haas et al. (2004). For the (regime–dependent) conditional distributions, we use the symmetric and the Fernández and Steel (1998) skewed versions of the Normal and Student–t distributions. Overall, this leads to sixteen models.

Our second contribution is to test the impact of the estimation method on the performance of the volatility forecasting model. GARCH and MSGARCH models are traditionally estimated with a frequentist (typically via ML) approach; see Haas et al. (2004), Marcucci (2005) and Augustyniak (2014). However, several recent studies have argued that a Bayesian approach offers some advantages. For instance, Markov chain Monte Carlo (MCMC) procedures can explore the joint posterior distribution of the model parameters, and parameter uncertainty is naturally integrated into the risk forecasts via the predictive distribution (Ardia, 2008; Bauwens et al., 2010, 2014; Geweke and Amisano, 2010; Ardia et al., 2017c).

Combining the sixteen model specifications with the frequentist and Bayesian estimation methods, we obtain 32 possible candidates for the state–of–the–art methodology for monitoring financial risk. We use an out–of–sample evaluation period of 2,000 days, that ranges from (approximately) 2005 to 2016 and consists of daily log–returns. We evaluate the accuracy of the risk prediction models in terms of estimating the Value–at–Risk (VaR), the Expected Shortfall (ES), and the left–tail (*i.e.*, losses) of the conditional distribution of the assets' returns.

Our empirical results suggest a number of practical insights which can be summarized as follows. First, we find that MSGARCH models report better VaR, ES, and left–tail distribution forecasts than their single–regime counterpart. This is especially true for stock return data. Moreover, improvements are more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model. Second, accounting for parameter uncertainty improves the accuracy of the left–tail predictions, independently of the inclusion of the Markov–switching mechanism. Moreover, larger improvements are observed in the case of single–regime models. Overall, we recommend risk managers to rely on more flexible models and to perform inference accounting for parameter uncertainty.

In addition to showing the good performance of MSGARCH models and Bayesian estimation methods, we refer risk managers to our R package MSGARCH (Ardia et al., 2017a,b), which implements MSGARCH models in the R statistical language with efficient C++ code.[3] We hope that this paper and the accompanying package will encourage practitioners and academics in the financial community to use MSGARCH models and Bayesian estimation methods.

The paper proceeds as follows. Model specification, estimation, and forecasting are presented in Section 2. The datasets, the testing design, and the empirical results are discussed in Section 3. Section 4 concludes.

2. Risk forecasting with Markov–switching GARCH models

A key aspect in quantitative risk management is the modeling of the risk drivers of the securities held by the fund manager. We consider here the univariate parametric framework, that computes the desired risk measure in four steps. First, a statistical model which describes the daily log–returns (profit and loss, P&L) dynamics is determined. Second, the model parameters are estimated for a given estimation window. Third, the one/multi–day ahead distribution of log–returns is obtained (either analytically or by simulation). Fourth, relevant risk measures such as the Value–at–Risk (VaR) and the Expected Shortfall (ES) are computed from the distribution. The VaR represents a quantile of the distribution of log–returns at the desired horizon, and the ES is the expected loss when the loss exceeds the VaR level (Jorion, 2006). Risk managers can then allocate

[3] Our research project was funded by the 2014 SAS/IIF forecasting research grant, to compare MSGARCH vs. GARCH models, and to develop and render publicly available the computer code for the estimation of MSGARCH models.

risk capital given their density or risk measure forecasts. Also, they can assess the quality of the risk model, *ex–post*, via statistical procedures referred to as *backtesting*.

2.1. Model specification

We define $y_t \in \mathbb{R}$ as the (percentage point) log–return of a financial asset at time t. To simplify the exposition, we assume that the log–returns have zero mean and are not autocorrelated.[4] The general Markov–switching GARCH specification can be expressed as:

$$y_t \,|\, (s_t = k, \mathcal{I}_{t-1}) \sim \mathcal{D}(0, h_{k,t}, \boldsymbol{\xi}_k) \,, \tag{1}$$

where $\mathcal{D}(0, h_{k,t}, \boldsymbol{\xi}_k)$ is a continuous distribution with zero mean, time–varying variance $h_{k,t}$, and additional shape parameters (*e.g.*, asymmetry) gathered in the vector $\boldsymbol{\xi}_k$.[5] Furthermore, we assume that the latent variable s_t, defined on the discrete space $\{1, \ldots, K\}$, evolves according to an unobserved first order ergodic homogeneous Markov chain with transition probability matrix $\mathbf{P} \equiv \{p_{i,j}\}_{i,j=1}^{K}$, with $p_{i,j} \equiv \mathbb{P}[s_t = j \,|\, s_{t-1} = i]$. We denote by $\mathcal{I}_{t-1}$ the information set up to time $t-1$, that is, $\mathcal{I}_{t-1} \equiv \{y_{t-i}, i > 0\}$. Given the parametrization of $\mathcal{D}(\cdot)$, we have $\mathbb{E}[y_t^2 \,|\, s_t = k, \mathcal{I}_{t-1}] = h_{k,t}$, that is, $h_{k,t}$ is the variance of y_t conditional on the realization of s_t and the information set $\mathcal{I}_{t-1}$.

As in Haas et al. (2004), the conditional variance of y_t is assumed to follow a GARCH–type model. More precisely, conditionally on regime $s_t = k$, $h_{k,t}$ is specified as a function of past returns and the additional regime–dependent vector of parameters $\boldsymbol{\theta}_k$:

$$h_{k,t} \equiv h(y_{t-1}, h_{k,t-1}, \boldsymbol{\theta}_k) \,,$$

where $h(\cdot)$ is a $\mathcal{I}_{t-1}$–measurable function, which defines the filter for the conditional variance and also ensures its positiveness. We further assume that $h_{k,1} \equiv \bar{h}_k$ ($k = 1, \ldots, K$), where $\bar{h}_k$ is a fixed initial variance level for regime k, that we set equal to the unconditional variance in regime k.

[4]In practice, this means that we apply the (MS)GARCH models to de–meaned log–returns, as explained in Section 3.

[5]For $t = 1$, we initialize the regime probabilities and the conditional variances at their unconditional levels. To simplify exposition, we use henceforth for $t = 1$ the same notation as for general t, since there is no confusion possible.

Depending on the form of $h(\cdot)$, we obtain different scedastic specifications. For instance, if:

$$h_{k,t} \equiv \omega_k + \alpha_k y_{t-1}^2 + \beta_k h_{k,t-1} \,,$$

with $\omega_k > 0$, $\alpha_k > 0$, $\beta_k \geq 0$ and $\alpha_k + \beta_k < 1$ ($k = 1, \ldots, K$), we obtain the Markov–switching GARCH$(1,1)$ model presented in Haas et al. (2004).[6] In this case $\boldsymbol{\theta}_k \equiv (\omega_k, \alpha_k, \beta_k)'$.

Alternative definitions of the function $h(\cdot)$ can be easily incorporated in the model. For instance, to account for the well–known asymmetric reaction of volatility to the sign of past returns (often referred to as the *leverage effect*; see Black 1976), we specify a Markov–switching GJR$(1,1)$ model exploiting the volatility specification of Glosten et al. (1993):

$$h_{k,t} \equiv \omega_k + (\alpha_k + \gamma_k \mathbb{I}\{y_{t-1} < 0\})\, y_{t-1}^2 + \beta_k h_{k,t-1} \,,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, that is equal to one if the condition holds, and zero otherwise. In this case, the additional parameter $\gamma_k \geq 0$ controls the asymmetry in the conditional variance process. We have $\boldsymbol{\theta}_k \equiv (\omega_k, \alpha_k, \gamma_k, \beta_k)'$. Covariance–stationarity of the variance process conditionally on the Markovian state is achieved by imposing $\alpha_k + \beta_k + \kappa_k \gamma_k < 1$, where $\kappa_k \equiv \mathbb{P}[y_t < 0 \,|\, s_t = k, \mathcal{I}_{t-1}]$. For symmetric distributions we have $\kappa_k = 1/2$. For skewed distributions, κ_k is obtained following the approach of Trottier and Ardia (2016).

We consider different choices for $\mathcal{D}(\cdot)$. We take the standard Normal ($\mathcal{N}$) and the Student–t ($\mathcal{S}$) distributions. To investigate the benefits of incorporating skewness in our analysis, we also consider the standardized skewed version of $\mathcal{N}$ and $\mathcal{S}$ obtained using the mechanism of Fernández and Steel (1998) and Bauwens and Laurent (2005); see Trottier and Ardia (2016) for more details. We denote the standardized skew–Normal and the skew–Student–t by sk$\mathcal{N}$ and sk$\mathcal{S}$, respectively.

Overall, our model set includes 16 different specifications recovered as combinations of:

- The number of regimes, $K \in \{1, 2\}$. When $K = 1$, we label our specification as single–regime (SR), and, when $K = 2$, as Markov–switching (MS);

[6] We require that the conditional variance in each regime is covariance–stationary. This is a stronger condition than in Haas et al. (2004), but this allows us to ensure stationarity for various forms of conditional variance and/or conditional distributions.

- The conditional variance specification: GARCH$(1,1)$ and GJR$(1,1)$;
- The choice of the conditional distribution $\mathcal{D}(\cdot)$, that is, $\mathcal{D} \in \{\mathcal{N}, \mathcal{S}, \mathrm{sk}\mathcal{N}, \mathrm{sk}\mathcal{S}\}$.[7]

2.2. Estimation

We estimate the models either through frequentist or Bayesian techniques. Both approaches require the evaluation of the likelihood function.

In order to write the likelihood function corresponding to the MSGARCH model specification (1), we regroup the model parameters into $\boldsymbol{\Psi} \equiv (\boldsymbol{\xi}_1, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\xi}_K, \boldsymbol{\theta}_K, \mathbf{P})$. The conditional density of y_t in state $s_t = k$ given $\boldsymbol{\Psi}$ and $\mathcal{I}_{t-1}$ is denoted by $f_{\mathcal{D}}(y_t \,|\, s_t = k, \boldsymbol{\Psi}, \mathcal{I}_{t-1})$.

By integrating out the state variable s_t, we obtain the density of y_t given $\boldsymbol{\Psi}$ and $\mathcal{I}_{t-1}$ only. The (discrete) integration is obtained as follows:

$$f(y_t \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1}) \equiv \sum_{i=1}^{K} \sum_{j=1}^{K} p_{i,j}\, \eta_{i,t-1}\, f_{\mathcal{D}}(y_t \,|\, s_t = j, \boldsymbol{\Psi}, \mathcal{I}_{t-1}) \,, \tag{2}$$

where $\eta_{i,t-1} \equiv \mathbb{P}[s_{t-1} = i \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1}]$ is the filtered probability of state i at time $t-1$ and where we recall that $p_{i,j}$ denotes the transition probability of moving from state i to state j. The filtered probabilities $\{\eta_{k,t}; k = 1, \ldots, K; t = 1, \ldots, T\}$ are obtained via the Hamilton filter; see Hamilton (1989) and Hamilton (1994, Chapter 22) for details.

Finally, the likelihood function is obtained from (2) as follows:

$$\mathcal{L}(\boldsymbol{\Psi} \,|\, \mathcal{I}_T) \equiv \prod_{t=1}^{T} f(y_t \,|\, \boldsymbol{\Psi}, \mathcal{I}_{t-1}) \,. \tag{3}$$

The ML estimator $\widehat{\boldsymbol{\Psi}}$ is obtained by maximizing the logarithm of (3). In the case of the Bayesian estimation, the likelihood function is combined with a prior $f(\boldsymbol{\Psi})$ to build the kernel of the posterior

[7] We also tested the asymmetric EGARCH scedastic specification (Nelson, 1991) as well as alternative fat–tailed distributions, such as the Laplace and GED distributions. The performance results were qualitatively similar.

distribution $f(\boldsymbol{\Psi} \,|\, \mathcal{I}_T)$.[8] As the posterior is of an unknown form (the normalizing constant is numerically intractable), it must be approximated by simulation techniques. In our case, MCMC draws from the posterior are generated with the adaptive random–walk Metropolis sampler of Vihola (2012). We use 50,000 burn–in draws and build the posterior sample of size 1,000 with the next 50,000 draws keeping only every 50th draw to diminish the autocorrelation in the chain.[9] For both the frequentist and the Bayesian estimation, we ensure positivity and stationarity of the conditional variance in each regime during the estimation. Moreover, we impose constraints on the parameters to ensure that volatilities under the MSGARCH specification cannot be generated by a single–regime specification. In the case of the frequentist estimation, these constraints are enforced in the likelihood optimization by using mapping functions. For the Bayesian estimation, this is achieved through the prior.

2.3. Density and downside risk forecasting

Generating one–step ahead density and downside risk forecasts (VaR and ES) with MSGARCH models is straightforward. First, note that the one–step ahead conditional probability density function (PDF) of y_{T+1} is a mixture of K regime–dependent distributions:

$$f(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) \equiv \sum_{k=1}^{K} \pi_{k,T+1} f_{\mathcal{D}}(y_{T+1} \,|\, s_{T+1} = k, \boldsymbol{\Psi}, \mathcal{I}_T)\,, \tag{4}$$

[8]We build our prior from diffuse independent priors as follows:

$$\begin{aligned}
f(\boldsymbol{\Psi}) &\propto f(\boldsymbol{\theta}_1, \boldsymbol{\xi}_1) \cdots f(\boldsymbol{\theta}_K, \boldsymbol{\xi}_K) f(\mathbf{P})\, \mathbb{I}\{\bar{h}_1 < \cdots < \bar{h}_K\} \\
f(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k) &\propto f(\boldsymbol{\theta}_k) f(\boldsymbol{\xi}_k)\, \mathbb{I}\{(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k) \in \mathcal{CSC}_k\} \quad (k = 1, \ldots, K) \\
f(\boldsymbol{\theta}_k) &\propto f_{\mathcal{N}}(\boldsymbol{\theta}_k; \mathbf{0}, 1{,}000 \times \mathbf{I})\, \mathbb{I}\{\boldsymbol{\theta}_k > \mathbf{0}\} \quad (k = 1, \ldots, K) \\
f(\boldsymbol{\xi}_k) &\propto f_{\mathcal{N}}(\boldsymbol{\xi}_k; \mathbf{0}, 1{,}000 \times \mathbf{I})\, \mathbb{I}\{\xi_{k,1} > 0, \xi_{k,2} > 2\} \quad (k = 1, \ldots, K) \\
f(\mathbf{P}) &\propto \prod_{i=1}^{K} \left(\prod_{j=1}^{K} p_{i,j} \right) \mathbb{I}\{0 < p_{i,i} < 1\}\,,
\end{aligned}$$

where $\mathbf{0}$ and $\mathbf{I}$ denote a vector of zeros and an identity matrix of appropriate sizes, $f_{\mathcal{N}}(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\xi_{k,1}$ is the asymmetry parameter, and $\xi_{k,2}$ the tail parameter of the skewed Student–t distribution in regime k. Moreover, $\bar{h}_k \equiv \bar{h}_k(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ is the unconditional variance in regime k and $\mathcal{CSC}_k$ denotes the covariance–stationarity condition in regime k; see Trottier and Ardia (2016).

[9]We performed several sensitivity analyses to assess the impact of the estimation's setup. First, we changed the hyper–parameter values. Second, we ran longer MCMC chains. Third, we used 10,000 posterior draws instead of 1,000. Finally, we tested an alternative MCMC sampler based on adaptive mixtures of Student–t distribution (Ardia et al., 2009). In all cases, the conclusions remained qualitatively similar.

with mixing weights $\pi_{k,T+1} \equiv \sum_{i=1}^{K} p_{i,k}\eta_{i,T}$ where $\eta_{i,T} \equiv \mathbb{P}[s_T = i \,|\, \boldsymbol{\Psi}, \mathcal{I}_T]$ $(i = 1, \ldots, K)$ are the filtered probabilities at time T. The cumulative density function (CDF) is obtained from (4) as follows:

$$F(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) \equiv \int_{-\infty}^{y_{T+1}} f(z \,|\, \boldsymbol{\Psi}, \mathcal{I}_T)\mathrm{d}z \,. \tag{5}$$

Within the frequentist framework, the predictive PDF and CDF are simply computed by replacing $\boldsymbol{\Psi}$ by the ML estimator $\widehat{\boldsymbol{\Psi}}$ in (4) and (5). Within the Bayesian framework, we proceed differently, and integrate out the parameter uncertainty. Given a posterior sample $\{\boldsymbol{\Psi}^{[m]}, m = 1, \ldots, M\}$, the predictive PDF is obtained as:

$$f(y_{T+1} \,|\, \mathcal{I}_T) \equiv \int_{\boldsymbol{\Psi}} f(y_{T+1} \,|\, \boldsymbol{\Psi}, \mathcal{I}_T) f(\boldsymbol{\Psi} \,|\, \mathcal{I}_T)\mathrm{d}\boldsymbol{\Psi} \approx \frac{1}{M} \sum_{m=1}^{M} f(y_{T+1} \,|\, \boldsymbol{\Psi}^{[m]}, \mathcal{I}_T) \,. \tag{6}$$

The predictive CDF is given by:

$$F(y_{T+1} \,|\, \mathcal{I}_T) \equiv \int_{-\infty}^{y_{T+1}} f(z \,|\, \mathcal{I}_T)\mathrm{d}z \,. \tag{7}$$

For both estimation approaches, the VaR is estimated as a quantile of the predictive density, by numerically inverting the predictive CDF. For instance, in the Bayesian framework, the VaR at the α risk level equals:

$$\mathrm{VaR}_{T+1}^{\alpha} \equiv \inf \left\{ y_{T+1} \in \mathbb{R} \,|\, F(y_{T+1} \,|\, \mathcal{I}_T) = \alpha \right\} \,, \tag{8}$$

while the ES at the α risk level is given by:

$$\mathrm{ES}_{T+1}^{\alpha} \equiv \frac{1}{\alpha} \int_{-\infty}^{\mathrm{VaR}_{T+1}^{\alpha}} z f(z | \mathcal{I}_T)\mathrm{d}z \,. \tag{9}$$

In our empirical application, we consider the VaR and the ES at the 1% and 5% risk levels.

For evaluating the risk at an h–period horizon, we must rely on simulation techniques to obtain the conditional density and downside risk measures, as described, for instance, in Blasques et al. (2016). More specifically, given a MSGARCH model parameter $\boldsymbol{\Psi}$, we generate 25,000 paths of daily log–returns over a horizon of h days.[10] The simulated distribution and the obtained α–quantile then serve as estimates of the density and downside risk forecasts of the h–day cumulative log–return.

3. Large–scale empirical study

We use 1,500 log–returns (in percent) for the estimation and run the backtest over 2,000 out–of–sample log-returns for a period ranging from October 10, 2008, to November 17, 2016 (the full dataset starts on December 26, 2002). Each model is estimated on a rolling window basis, and one–step ahead as well as multi–step cumulative log–returns density forecasts are obtained.[11] From the estimated density, we compute the VaR and the ES at the 1% and 5% risk levels.

3.1. Datasets

We test the performance of the various models on several universes of securities typically traded by fund managers:

- A set of 426 stocks, selected by taking the S&P 500 universe index as of November 2016, and omitting the stocks for which more than 5% of the daily returns are zero, and stocks for which there are less than 3,500 daily return observations.
- A set of eleven stock market indices: (1) S&P 500 (US; SPX), (2) FTSE 100 (UK; FTSE), (3) CAC 40 (France; FCHI), (4) DAX 30 (Germany; GDAXI), (5) Nikkei 225 (Japan; N225), (6) Hang Seng (China, HSI), (7) Dow Jones Industrial Average (US; DJI), (8) Euro Stoxx 50 (Europe; STOXX50), (9) KOSPI (South Korea; KS11), (10) S&P/TSX Composite (Canada; GSPTSE), and (11) Swiss Market Index (Switzerland; SSMI);

[10] With the frequentist estimation, we generate 25,000 paths with parameter $\widehat{\mathbf{\Psi}}$, while in the case of the Bayesian estimation, we generate 25 paths for each of the 1,000 value $\mathbf{\Psi}^{[m]}$ in the posterior sample. We use this number to get enough draws from the predictive distribution as we focus on the left tail. Geweke (1989) shows that the consistent estimation of the predictive distribution does not depend on the number of paths generated from the posterior. So with 25 paths, we indeed converge to the correct predictive distribution. We verified that increasing the number of simulations has no material impact on the results.

[11] Model parameters are updated every ten observations. We selected this frequency to speed up the computations. Similar results for a subset of stocks were obtained when updating the parameters every day. This is also in line with the observation of Ardia and Hoogerheide (2014), who show, in the context of GARCH models, that the performance of VaR forecasts is not significantly affected when moving from a daily updating frequency to a weekly or monthly updating frequency. Note that while parameters are updated every ten observations, the density and downsides risk measures are computed every day.

- A set of eight foreign exchange rates: USD against CAD, DKK, NOK, AUD, CHF, GBP, JPY, and EUR.[12]

Data are retrieved from Datastream. Each price series is expressed in local currency. We compute the daily percentage log–return series defined by $x_t \equiv 100 \times \log(P_t/P_{t-1})$, where P_t is the adjusted closing price (value) on day t. We then de–mean the returns x_t using an AR(1)–filter, and use those filtered returns, y_t, to estimate and evaluate the precision of the financial risk monitoring systems.

In Table 1, we report the summary statistics on the out–of–sample daily, five–day, and ten–day cumulative log–returns for the three asset classes. We report the standard deviation (Std), the skewness (Skew) and kurtosis (Kurt) coefficients evaluated over the full sample as well as the historical 1% and 5% VaR and ES levels. We note the higher volatility in all periods for the universe of stocks, followed by indices and exchange rates. All securities exhibit negative skewness, with larger values for indices and stocks, while exchange rates seem to behave more symmetrically. Interestingly, the negative skewness tends to be more pronounced for indices as the horizon grows. Finally, at the daily horizon, we observe a significant kurtosis for stocks. Fat tails are also present for indices and exchange rates, but less pronounced than for stocks. However, as the horizon grows, the kurtosis of all asset classes tends to diminish.

[Insert Table 1 about here.]

3.2. Forecasting performance tests

We compare the adequacy of the 32 models in terms of providing accurate forecasts of the left tail of the conditional distribution and the VaR and ES levels.

3.2.1. Accuracy of VaR predictions

For testing the accuracy of the VaR predictions, we use the so–called *hit* variable, which is a dummy variable indicating a loss that exceeds the VaR level:

$$I_t^\alpha \equiv \mathbb{I}\{y_t \leq \mathrm{VaR}_t^\alpha\}\,,$$

[12] In the context of foreign exchange rates, left–tail forecasts aim at assessing the risk for a foreign investor investing in USD and therefore facing devaluation of USD.

where VaR_t^α denotes the VaR prediction at risk level α for time t, and $\mathbb{I}\{\cdot\}$ is the indicator function equal to one if the condition holds, and zero otherwise. If the VaR is correctly specified, then the hit variable has a mean value of α and is independently distributed over time. We test this for the $\alpha = 1\%$ and $\alpha = 5\%$ risk levels using the unconditional coverage (UC) test by Kupiec (1995), and the dynamic quantile (DQ) test by Engle and Manganelli (2004).

The UC test by Kupiec (1995) uses the likelihood ratio to test that the violations have a Binomial distribution with $\mathbb{E}[I_t^\alpha] = \alpha$. Denote by $x \equiv \sum_{t=1}^T I_t^\alpha$ the number of observed rejections on a total of T observations, then, under the null of correct coverage, we have that the test statistic:

$$\text{UC}_\alpha \equiv -2\ln\left[(1-\alpha)^{T-x}\alpha^x\right] + 2\ln\left[\left(1-\frac{x}{T}\right)^{T-x}\left(\frac{x}{T}\right)^x\right],$$

is asymptotically chi–square distributed with one degree–of–freedom.

The DQ test by Engle and Manganelli (2004) is a test of the joint hypothesis that $\mathbb{E}[I_t^\alpha] = \alpha$ and that the hit variables are independently distributed. The implementation of the test involves the de–meaned process $\text{Hit}_t^\alpha \equiv I_t^\alpha - \alpha$. Under correct model specification, unconditionally and conditionally, Hit_t^α has zero mean and is serially uncorrelated. The DQ test is then the traditional Wald test of the joint nullity of all coefficients in the following linear regression:

$$\text{Hit}_t^\alpha = \delta_0 + \sum_{l=1}^L \delta_l \text{Hit}_{t-l}^\alpha + \delta_{L+1}\text{VaR}_{t-1}^\alpha + \epsilon_t\,.$$

If we denote the OLS parameter estimates as $\widehat{\boldsymbol{\delta}} \equiv (\widehat{\delta}_0, \ldots, \widehat{\delta}_{L+1})'$ and $\mathbf{Z}$ as the corresponding data matrix with, in column, the observations for the $L+2$ explanatory variables, then the DQ test statistic of the null hypothesis of correct unconditional and conditional coverage is:

$$\text{DQ}_\alpha \equiv \frac{\widehat{\boldsymbol{\delta}}'\mathbf{Z}'\mathbf{Z}\widehat{\boldsymbol{\delta}}}{\alpha(1-\alpha)}\,.$$

As in Engle and Manganelli (2004), we choose $L = 4$ lags. Under the null hypothesis of correct unconditional and conditional coverage, we have that DQ_α is asymptotically chi–square distributed with $L+2$ degrees of freedom.[13]

3.2.2. Accuracy of the left–tail distribution

Risk managers care not only about the accuracy of the VaR forecasts but also about the accuracy of the complete left–tail region of the log–return distribution. This broader view of all losses is central in modern risk management, and, consistent with the regulatory shift to using Expected Shortfall as the risk measure for determining capital requirements starting in 2018 (Basel Committee on Banking Supervision, 2013). We evaluate the effectiveness of MSGARCH models to yield accurate predictions of the left–tail distribution in three ways.

A first approach is to compute the weighted average difference of the observed returns with respect to the VaR value, and give higher weight to losses that violate the VaR level. This corresponds to the quantile loss assessment of González-Rivera et al. (2004) and McAleer and Da Veiga (2008). Formally, given a VaR prediction at risk level α for time t, the associated quantile loss (QL) is defined as:

$$\mathrm{QL}_t^\alpha \equiv (\alpha - I_t^\alpha)(y_t - \mathrm{VaR}_t^\alpha)\,.$$

The choice of this loss function for VaR assessment is appropriate since quantiles are elicited by it; that is, when the conditional distribution is static over the sample, the VaR_t^α can be estimated by minimizing the average quantile loss function. Elicitability is useful for model selection, estimation, forecast comparison, and forecast ranking.

Unfortunately, there is no loss function available for which the ES risk measure is elicitable; see, for instance, Bellini and Bignozzi (2015) and Ziegel (2016). However, it has been recently shown by Fissler and Ziegel (2016) (FZ) that, in case of a constant conditional distribution, the couple (VaR, ES) is jointly elicitable, as the values of v_t and e_t that minimize the sample average of the following loss function:

$$FZ(y_t, v_t, e_t, \alpha, G_1, G_2) \equiv (I_t^\alpha - \alpha)\left(G_1(v_t) - G_1(y_t) + \frac{1}{\alpha}G_2(e_t)v_t\right) - G_2(e_t)\left(\frac{1}{\alpha}I_t^\alpha y_t - e_t\right) - \mathcal{G}_2(e_t)\,,$$

[13] As in Bams et al. (2017), it is possible to add more explanatory variable such as lagged returns and lagged squared returns and jointly test the new coefficients. In our case, results obtained by adding lagged returns or lagged squared returns are qualitatively similar to the simpler specification.

where G_1 is weakly increasing, G_2 is strictly positive and strictly increasing, and $\mathcal{G}_2' = G_2$. In a similar setup as ours, Patton et al. (2017) assume the values of VaR and ES to be strictly negative and recommend setting $G_1(x) = 0$ and $G_2(x) = -1/x$. For a VaR and a ES prediction at risk level α for time t, the associated joint loss function (FZL) is then given by:

$$\mathrm{FZL}_t^\alpha \equiv \frac{1}{\alpha\,\mathrm{ES}_t^\alpha}\, I_t^\alpha\,(y_t - \mathrm{VaR}_t^\alpha) + \frac{\mathrm{VaR}_t^\alpha}{\mathrm{ES}_t^\alpha} + \log(-\mathrm{ES}_t^\alpha) - 1\,, \tag{10}$$

for $\mathrm{ES}_t^\alpha \leq \mathrm{VaR}_t^\alpha < 0$. Hence, in order to gauge the precision of both the VaR and ES downside risk estimates, we use the FZL function as our second evaluation criterion.

A third approach that we consider is to compare the empirical distribution with the predicted conditional distribution through the weighed Continuous Ranked Probability Score (wCRPS), introduced by Gneiting and Ranjan (2011) as a generalization of the CRPS scoring rule (Matheson and Winkler, 1976). Following the notation introduced in Section 2, the wCRPS for a forecast at time t is defined as:

$$\mathrm{wCRPS}_t \equiv \int_{\mathbb{R}} \omega(z)\,\left(F(z\,|\,\mathcal{I}_{t-1}) - \mathbb{I}\{y_t \leq z\}\right)^2 \mathrm{d}z\,,$$

where F is the predictive CDF and $\omega\colon \mathbb{R} \to \mathbb{R}^+$ is a continuous weight function, which emphasizes regions of interest of the predictive distribution, such as the tails or the center. Since our focus is on predicting losses, we follow Gneiting and Ranjan (2011) and use the decreasing weight function $\omega(z) \equiv 1 - \Phi(z)$, where Φ is the CDF of a standard Gaussian distribution. This way, discrepancies in the left tail of the return distribution are weighed more than those in the right tail.[14]

For the QL, FZL and wCRPS approaches, we test the statistical significance of the differences in the forecasting performance of two competing models, say models i and j. We do this by first

[14]We follow the implementation of Gneiting and Ranjan (2011) and compute wCRPS with the following approximation:

$$\mathrm{wCRPS}_t \approx \frac{z_u - z_l}{M-1} \sum_{m=1}^{M} w(z_m)\,\left(F(z_m\,|\,\mathcal{I}_{t-1}) - \mathbb{I}\{y_t \leq z_m\}\right)^2\,,$$

where $z_m \equiv z_l + m \times (z_u - z_l)/M$ and z_u and z_l are the upper and lower values, which defines the range of integration. The accuracy of the approximation can be increased to any desired level by M. Setting $z_l = -100$, $z_u = 100$ and $M = 1{,}000$ provides an accurate approximation when working with returns in percentage points. We also tested the triangular integration approach and results were numerically equivalent. Alternative weights specifications, focusing on the right tail, center, of full distribution, lead to similar conclusions at the one–day forecasting horizon. The results are available from the authors upon request.

computing, for each out–of–sample date t, the average performance statistics across all securities in the same asset class. Denote this difference as $\Delta_t^{i-j} \equiv L_t^i - L_t^j$, where L_t^i is the average value of the performance measure (QL, FZL or wCRPS) of all assets within the same asset class. We then test $H_0 : \mathbb{E}[\Delta_t^{i-j}] = 0$ using the standard Diebold and Mariano (1995) (DM) test statistic, implemented with the heteroscedasticity and autocorrelation robust (HAC) standard error estimators of Andrews (1991) and Andrews and Monahan (1992). If the null hypothesis is rejected, the sign of the test statistics indicates which model is, on average, preferred for a particular loss measure.

3.3. Results

We now summarize the results regarding our main research question: *Does the additional complexity of Markov–switching and the use of Bayesian estimation methods lead to more accurate out–of–sample downside risk predictions?* We first present our results regarding the accuracy of the VaR predictions and then use the QL, FZL and wCRPS approaches to evaluate the gains in terms of left–tail predictions.

3.3.1. Effect of model and estimator choice on the accuracy of VaR predictions

We first use the UC test of Kupiec (1995) and the DQ test of Engle and Manganelli (2004) to evaluate the accuracy of each of the 32 methods considered in terms of predicting the VaR at the 5% and 1% level for the daily returns on the 426 stocks, 11 stock indices and 8 exchange rates. For each asset, we obtain the p–value corresponding to the UC and DQ test computed using 2,000 out–of–sample observations. In Table 2, we aggregate the results per asset class by presenting the percentage of assets for which the null hypothesis of correct unconditional and conditional coverage is rejected at the 5% level, by the UC and DQ test, respectively.[15]

[Insert Table 2 about here.]

[15] In the case of stocks, as the universe is large and therefore prone to false positives, the p–values are corrected for Type I error using the false discovery rate (FDR) approach of Benjamini and Hochberg (1995). The FDR correction for a confidence level q proceeds as follows. For a set of m ordered p–values $p_1 \leq p_2 \leq \ldots \leq p_m$ and corresponding null hypotheses $H_1, H_2, \ldots, H_m$, define v as the largest value of i for which $p_i \leq \frac{i}{m} q$, and the reject all hypotheses H_i for $i = 1, \ldots, v$.

Consider in Panels A and B of Table 2 the results for the UC test. At both VaR risk levels, we find that the validity of the VaR predictions based on the GARCH and GJR skewed Student–t risk model is never rejected, whatever the use of SR or MS models, or frequentist or Bayesian estimation methods. The result changes drastically when we consider the more powerful DQ test of correct conditional coverage in Panels C and D. Here, we find clear evidence that the use of MS GJR models leads to a lower percentage of rejections of the validity of the VaR prediction for all asset classes. At the 1% risk level, these differences are most often significant.

Overall, the one–day ahead backtest results indicate outperformance of MS over SR models, especially for VaR prediction on equities. Moreover, a GJR specification leads to a substantial reduction in the rejection frequencies. Both for MS and SR specifications, a fat–tailed conditional distribution is of primary importance and delivers excellent results at both risk levels.

Finally, for this analysis, the frequency of rejections are similar between the Bayesian and frequentist estimation methods. More precisely, a t–test for equal average rejections indicates that differences are insignificant. We thus conclude that, based on the analysis of VaR forecast accuracy, it is hard to discriminate between the estimation methods.

3.3.2. Effect of model choice on accuracy of left–tail predictions

A further question is how model simplification affects the accuracy of the left–tail return prediction. In Table 3, we report the standardized difference between the average QL, FZL and wCRPS values of the assets belonging to the same asset class, when we switch from a MS specification to a SR specification. The standardization corresponds to the Diebold and Mariano (1995) (DM) test statistic. Negative values indicate out–of–sample evidence of a deterioration in the prediction accuracy when using the SR specification instead of the MS specification. When the standardized value exceeds 2.57 (*i.e.*, the critical value computed using a 1% significance level for a bilateral test based on the asymptotic Normal distribution) in absolute value, the statistical significance is highlighted with a gray shading.[16] We report results obtained with the Bayesian framework only,

[16]We take the standard critical value in Diebold and Mariano (1995) as our Markov–switching specifications do not nest the alternative single–regime model due to parameter constraints imposing that the volatility dynamics are numerically different in each regime, and that each regime has a non–zero probability. The approach by Clark and McCracken (2001) should be used when comparing nested models.

as the performance obtained with the Bayesian estimation is better for both MS and SR models (especially for SR specifications) compared with the frequentist estimation.[17]

[Insert Table 3 about here.]

One–step ahead results for wCRPS favor MS models with negative values observed for almost all asset classes and model specifications. QL, FZL and wCRPS results are consistent with the backtest results: They confirm the superior performance of the MS specification for the universe of stocks, while outperformance is less clear for indices and exchange rates. Indeed, for indices, MS is required only when a non fat–tailed conditional distribution is assumed, while for exchange rates, MS is generally not required. Note that, for all assets, the improvements tend to be more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model.

For stocks, the MS specification significantly outperforms in terms of the FZL and wCRPS measures at the five–day horizon. For the wCRPS measure at the ten–day horizon, and for the QL measure at the five– and ten–day horizons, results are mostly insignificant, except for the FZL 5% measure, which favors MS models when a non fat–tailed conditional distribution is assumed. MS and SR models perform similarly for the five– and ten–day returns on stock indices. Finally, for exchange rate returns, SR models outperform MS models at the five– and ten–day horizons according to the QL 1% measure, while the differences in QL 5%, FZL, and wCRPS are insignificant.

It is informative to examine if these gains in forecasting precision are stable across the out–of–sample window. To determine this, we display in Figure 1 the cumulative wCRPS average loss differential over the whole out–of–sample period for the best performing specification, the GJR skewed Student–t model. Interestingly, we find that MSGARCH systematically outperforms GARCH according to the criteria that are most sensitive to the extreme left tail of the return distribution, namely the FZL (for $\alpha = 1\%$ and $\alpha = 5\%$) and QL (for $\alpha = 1\%$). We also notice that in these cases the gains of MSGARCH over GARCH increase during the last phase of the turbulent period 2008–2012. With regards to wCRPS and QL at $\alpha = 5\%$, we find that MSGARCH

[17] Hence, our discussion based on Bayesian results is more conservative in the sense that it gives an advantage to the SR specifications.

starts outperforming GARCH after the end of the turbulent period 2008–2012. We conjecture that this improvement in performance can be explained by the lack of flexibility of the single–regime GARCH specification. As also evident from the first panel of Figure 1, the market volatility has changed both its unconditional level and its dependence structure between the two periods 2008–2012 and 2012–2015. Since the estimation window is of 1'500 observations (approximately 7 years), observations in the period 2008–2012 affect GARCH predictions for the whole 2012–2015 forecasting period. Differently, MSGARCH allows the volatility process to adapt more rapidly to changes in regimes, resulting in better risk predictions. This is the case for the first half of the window, ranging from December 2008 to November 2012 and encompasses the Great Financial Crisis, but as well for the half of the window, ranging from December 2012 to November 2016 and follows the crisis; more calm market period.

[Insert Figure 1 about here.]

We now consider in Table 4 a complete comparison of the wCRPS performance of all MS models (in row) versus all SR models (in column). The elements in the diagonal correspond to the wCRPS values reported in Table 3. They are informative about the change in wCRPS when switching from a MS model to a SR model, keeping the same specification for the conditional variance and distribution. The analysis of the extra–diagonal elements is informative about the changes in wCRPS when switching from a MS model to a SR model, and changing the specification of the volatility model or the density function. In this table, an outperforming MS risk model is a model for which all standardized gains when changing the specification are negative. For almost all comparisons, this is the case for the MS GJR model with skewed Student–t innovations. The only exception is for modeling the returns of stock market indices, where it performs similarly as its SR counterpart.

[Insert Table 4 about here.]

3.3.3. Effect of estimator choice on accuracy of left–tail predictions

In Table 5, we report the results for the Bayesian versus frequentist estimation methods in the case of one–step ahead QL, FZL and wCRPS measures. Panel A (Panel B) shows the results for

MS (SR) models, where a negative (positive) value indicates outperformance (underperformance) of Bayesian against frequentist estimation. In light gray, we emphasize cases of significant outperformance of the Bayesian estimation over the frequentist approach. For stocks, the QL 1% and 5% comparisons indicate that Bayesian is preferred over ML, and it is significant in the majority of the specifications. The same observation can be made when using the FZL and wCRPS evaluation criteria. For stock indices and exchange rates, QL, FZL and wCRPS results are in favor of the Bayesian estimation for both MS and SR models but results are less significant than for stocks. Overall, we recommend to account for parameter uncertainty especially for stocks data, and when the interest is on the left tail of the log–returns distribution. The performance gain is especially large for SR models.

[Insert Table 5 about here.]

3.3.4. Constrained Markov–switching specifications

So far, our empirical results have highlighted the need for a MS mechanism in GARCH–type models in the case of stocks. We now refine the analysis by examining whether the same gains are achieved when constraining that the conditional distribution of the MS specifications has the same shape parameter across the regimes. Hence, we apply the MS mechanism only to the conditional variance. The objective is to determine whether, in the context of MS models, the switches in the variance dynamics are the dominant contributor to the gains in risk forecasting accuracy.

In Table 6, we report the performance measures obtained with the constrained MS models for the various horizons, when models are estimated with the Bayesian approach.[18] Results are in line with the non–constrained case of Table 3, but less significant. Hence, accounting for structural breaks in only the variance dynamics improves the risk forecasts at the daily, weekly and ten–day horizons. If we let the shape parameters depend upon the regime, we further improve the performance.

[Insert Table 6 about here.]

[18] Forecasting results obtained via frequentist estimation are qualitatively similar and available from the authors upon request.

4. Conclusion

In this paper, we investigate if MSGARCH models provide risk managers with useful tools for improving the risk forecasts of securities typically hold by fund managers. Moreover, we investigate if integrating the model's parameter uncertainty within the forecasts, via the Bayesian approach, improves predictions. Our results and practical advice can be summarized as follows.

First, risk managers should extend their GARCH–type models with a Markov–switching specification in case of investment in equities. Indeed, we find that Markov–switching GARCH models report better Value–at–Risk, Expected Shortfall, and left–tail distribution forecasts than their single–regime counterpart. This is especially true for stock return data. Moreover, improvements are more pronounced when the Markov–switching mechanism is applied to simple specifications such as the GARCH–Normal model.

Second, accounting for parameter uncertainty helps for left–tail predictions independently of the inclusion of the Markov–switching mechanism. Moreover, larger improvements are observed when parameter uncertainty is included in single–regime models.

Overall, we recommend risk managers to rely on more flexible models and to perform inference accounting for parameter uncertainty. To help them implementing these in practice, we have released the open–source R package MSGARCH; see Ardia et al. (2017a,b).

Our research could be extended in several ways. First, our study considered single–regime versus two–state Markov–switching specifications. Hence, it would be of interest to see if a third regime leads to superior performance, and if the optimal number of regimes (according to penalized likelihood information criteria) changes over time and is different across data sets. Second, additional universes could be considered, such as emerging markets and commodities. Third, one could extend the set of models and compare the performance of MSGARCH with realized volatility models such as the HEAVY model of Shephard and Sheppard (2010). Fourth, as suggested by a referee, it would be interesting to shed light on the parameter configurations for which the MSGARCH predictions can be expected to yield the higher improvement in risk forecast precision. An exploratory analysis has shown that a high persistence of at least one state seems needed to have a substantial difference in precision between MSGARCH and single–regime GARCH downside risk forecasts. A definite answer to this question is beyond the scope of this paper. Finally, our analysis only considered financial risk monitoring systems for individual financial assets. The new standard for capital requirements for market risk (Basel Committee on Banking Supervision, 2016) calls for backtesting at the individual desk level and the aggregate level. For this reason, it would be interesting to consider also the impact of choices in modeling dependence. Including these extensions in our current research setup increases further the (already large) number of models included in the comparison. We leave them as a topic for future work.

References

Andrews, D., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858. doi:10.2307/2938229.

Andrews, D., Monahan, J., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 60, 953–966. doi:10.2307/2951574.

Ardia, D., 2008. Financial Risk Management with Bayesian Estimation of GARCH Models: Theory and Applications. Springer, Heidelberg. doi:10.1007/978-3-540-78657-3.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., Peterson, B., Trottier, D.A., 2017a. MSGARCH: Markov-switching GARCH models in R. URL: https://cran.r-project.org/package=MSGARCH.

Ardia, D., Bluteau, K., Boudt, K., Catania, L., Trottier, D.A., 2017b. Markov-switching GARCH models in R: The MSGARCH package. URL: https://ssrn.com/abstract=2845809. working paper.

Ardia, D., Hoogerheide, L.F., 2014. GARCH models for daily stock returns: Impact of estimation frequency on Value-at-Risk and Expected Shortfall forecasts. Economics Letters 123, 187–190. doi:10.1016/j.econlet.2014.02.008.

Ardia, D., Hoogerheide, L.F., van Dijk, H.K., 2009. Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package AdMit. Journal of Statistical Software 29, 1–32. doi:10.18637/jss.v029.i03.

Ardia, D., Kolly, J., Trottier, D.A., 2017c. The impact of parameter and model uncertainty on market risk predictions from GARCH-type models. Journal of Forecasting 36, 808–823. doi:10.1002/for.2472.

Augustyniak, M., 2014. Maximum likelihood estimation of the Markov-switching GARCH model. Computational Statistics & Data Analysis 76, 61–75. doi:10.1016/j.csda.2013.01.026.

Bams, D., Blanchard, G., Lehnert, T., 2017. Volatility measures and Value-at-Risk. International Journal of Forecasting 33, 848–863. doi:10.1016/j.ijforecast.2017.04.004.

Basel Committee on Banking Supervision, 2013. Fundamental review of the trading book: A revised market risk framework. Technical Report 265. Bank of International Settlements.

Basel Committee on Banking Supervision, 2016. Minimum capital requirements for market risk. techreport 352. Bank of International Settlements.

Bauwens, L., Backer, B.D., Dufays, A., 2014. A Bayesian method of change-point estimation with recurrent regimes: Application to GARCH models. Journal of Empirical Finance 29, 207–229. doi:10.1016/j.jempfin.2014.06.008.

Bauwens, L., Laurent, S., 2005. A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. Journal of Business & Economic Statistics 23, 346–354. doi:10.1198/073500104000000523.

Bauwens, L., Preminger, A., Rombouts, J.V.K., 2010. Theory and inference for a Markov switching GARCH model. Econometrics Journal 13, 218–244. doi:10.1111/j.1368-423X.2009.00307.x.

Bellini, F., Bignozzi, V., 2015. On elicitable risk measures. Quantitative Finance 15, 725–733.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B Vol. 57, 289–300. URL: http://www.jstor.org/stable/2346101.

Black, F., 1976. Studies of stock price volatility changes, in: Proceedings of the 1976 Meetings of the American Statistical Association, pp. 177–181.

Blasques, F., Koopman, S.J., Łasak, K., Lucas, A., 2016. In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. International Journal of Forecasting 32, 875–887. doi:10.1016/j.ijforecast.2015.11.018.

Board of Governors of the Federal Reserve Systems, 2012. 99th Annual Report. Technical Report. Board of Governors of the Federal Reserve Systems.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. doi:10.1016/0304-4076(86)90063-1.

Cai, J., 1994. A Markov model of switching-regime ARCH. Journal of Business & Economic Statistics 12, 309–316. doi:10.2307/1392087.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110. doi:10.1016/S0304-4076(01)00071-9.

Danielsson, J., 2011. Risk and crises. VoxEU.org URL: http://voxeu.org/article/risk-and-crises-how-models-failed-and-are-failing.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263. doi:10.1080/07350015.1995.10524599.

Dueker, M.J., 1997. Markov switching in GARCH processes and mean-reverting stock-market volatility. Journal of Business & Economic Statistics 15, 26–34. doi:10.2307/1392070.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1008. doi:10.2307/1912773.

Engle, R.F., 2004. Risk and volatility: Econometric models and financial practice. American Economic Review 94, 405–420. doi:10.1257/0002828041464597.

Engle, R.F., Manganelli, S., 2004. CAViaR: Conditional autoregressive Value at Risk by regression quantiles. Journal of Business & Economic Statistics 22, 367–381. doi:10.1198/073500104000000370.

Fernández, C., Steel, M.F.J., 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association 93, 359–371. doi:10.1080/01621459.1998.10474117.

Fissler, T., Ziegel, J.F., 2016. Higher order elicitability and Osband's principle. The Annals of Statistics 44, 1680–1707. doi:10.1214/16-AOS1439.

Geweke, J., 1989. Exact predictive densities for linear models with ARCH disturbances. Journal of Econometrics 40, 63–86. doi:10.1016/0304-4076(89)90030-4.

Geweke, J., Amisano, G., 2010. Comparing and evaluating Bayesian predictive distributions of asset returns. International Journal of Forecasting 26, 216–230. doi:10.1016/j.ijforecast.2009.10.007.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48, 1779–1801. doi:10.1111/j.1540-6261.1993.tb05128.x.

Gneiting, T., Ranjan, R., 2011. Comparing density forecasts using threshold –and quantile– weighted scoring rules. Journal of Business & Economic Statistics 29, 411–422. doi:10.1198/jbes.2010.08110.

González-Rivera, G., Lee, T.H., Mishra, S., 2004. Forecasting volatility: A reality check based on option pricing, utility function, Value-at-Risk, and predictive likelihood. International Journal of Forecasting 20, 629–645. doi:10.1016/j.ijforecast.2003.10.003.

Gray, S.F., 1996. Modeling the conditional distribution of interest rates as a regime-switching process. Journal of Financial Economics 42, 27–62. doi:10.1016/0304-405x(96)00875-6.

Haas, M., Mittnik, S., Paolella, M.S., 2004. A new approach to Markov-switching GARCH models. Journal of Financial Econometrics 2, 493–530. doi:10.1093/jjfinec/nbh020.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384. doi:10.2307/1912559.

Hamilton, J.D., 1994. Time Series Analysis. First ed., Princeton University Press, Princeton, USA.

Hamilton, J.D., Susmel, R., 1994. Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics 64, 307–333. doi:10.1016/0304-4076(94)90067-1.

Hansen, P.R., Huang, Z., Shek, H.H., 2011. Realized GARCH: A joint model for returns and realized measures of volatility. Journal of Applied Econometrics 27, 877–906. doi:10.1002/jae.1234.

Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics 20, 873–889. doi:10.1002/jae.800.

Herwartz, H., 2017. Stock return prediction under GARCH – An empirical assessment. International Journal of Forecasting 33, 569–580. doi:10.1016/j.ijforecast.2017.01.002.

Jacquier, E., Polson, N.G., Rossi, P.E., 1994. Bayesian analysis of stochastic volatility models. Journal of Business & Economic Statistics 12, 371–389. doi:10.1080/07350015.1994.10524553.

Jorion, P., 2006. Value at Risk – The New Benchmark for Managing Financial Risk. Third ed., McGraw–Hill.

Klaassen, F., 2002. Improving GARCH volatility forecasts with regime-switching GARCH, in: Advances in Markov-Switching Models. Springer-Verlag, pp. 223–254. doi:10.1007/978-3-642-51182-0_10.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives 3. doi:10.3905/jod.1995.407942.

Lamoureux, C.G., Lastrapes, W.D., 1990. Persistence in variance, structural change, and the GARCH model. Journal of Business & Economic Statistics 8, 225–234. doi:10.2307/1391985.

Marcucci, J., 2005. Forecasting stock market volatility with regime-switching GARCH models. Studies in Nonlinear Dynamics & Econometrics 9. doi:10.2202/1558-3708.1145.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Management Science 22, 1087–1096. doi:10.1287/mnsc.22.10.1087.

McAleer, M., Da Veiga, B., 2008. Single-index and portfolio models for forecasting Value-at-Risk thresholds. Journal of Forecasting 27, 217–235. doi:10.1002/for.1054.

McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools. Second ed., Princeton University Press.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59, 347–370. doi:10.2307/2938260.

Opschoor, A., van Dijk, D., van der Wel, M., 2017. Combining density forecasts using focused scoring rules. Journal of Applied Econometrics doi:10.1002/jae.2575. in press.

Patton, A.J., Ziegel, J.F., Chen, R., 2017. Dynamic semiparametric models for expected shortfall. URL: https://ssrn.com/abstract=3000465. working paper.

Shephard, N., Sheppard, K., 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 25, 197–231. doi:10.1002/jae.1158.

Taylor, S.J., 1994. Modeling stochastic volatility: A review and comparative study. Mathematical Finance 4, 183–204. doi:10.1111/j.1467-9965.1994.tb00057.x.

Trottier, D.A., Ardia, D., 2016. Moments of standardized Fernández-Steel skewed distributions: Applications to the estimation of GARCH-type models. Finance Research Letters 18, 311–316. doi:10.1016/j.frl.2016.05.006.

Vihola, M., 2012. Robust adaptive Metropolis algorithm with coerced acceptance rate. Statistics and Computing 22, 997–1008. doi:10.1007/s11222-011-9269-5.

Ziegel, J.F., 2016. Coherence and elicitability. Mathematical Finance 26, 901–918.

Table 1: Summary statistics of the return data

The table presents the summary statistics of the (de–meaned) h–day cumulative log–returns for securities in the three asset classes used in our study. We report the standard deviation (Std), the skewness (Skew), the kurtosis (Kurt), and the 5% and 1% historical VaR and ES, on an unconditional basis for the 2,000 out–of–sample observations. For each statistic, we compute the 25th, 50th and 75th percentiles over the whole universe of assets.

h	Percentile	Std	Skew	Kurt	1% VaR	5% VaR	1% ES	5% ES
Panel A: Stocks (426 series)								
	25th	1.48	−0.39	6.89	−6.55	−3.44	−9.30	−5.53
1	50th	1.89	−0.13	9.24	−5.23	−2.85	−7.31	−4.50
	75th	2.33	0.12	14.10	−4.10	−2.25	−5.68	−3.50
	25th	3.29	−0.42	4.93	−14.60	−7.94	−19.14	−12.11
5	50th	4.21	−0.20	5.87	−11.59	−6.55	−14.84	−9.82
	75th	5.19	0.01	7.53	−9.15	−5.17	−12.00	−7.71
	25th	4.54	−0.49	4.47	−19.99	−10.92	−25.42	−16.54
10	50th	5.76	−0.27	5.30	−15.74	−9.02	−20.28	−13.19
	75th	6.98	−0.05	6.92	−12.43	−7.16	−16.08	−10.46
Panel B: Stock market indices (11 series)								
	25th	1.07	−0.40	6.07	−3.70	−2.37	−4.84	−3.30
1	50th	1.15	−0.23	7.29	−3.39	−1.85	−4.31	−2.78
	75th	1.39	−0.17	10.29	−3.05	−1.77	−4.01	−2.58
	25th	2.42	−0.55	5.04	−8.38	−5.09	−10.65	−7.30
5	50th	2.54	−0.47	6.18	−7.60	−4.22	−9.85	−6.17
	75th	3.09	−0.29	8.22	−6.91	−3.86	−9.22	−5.97
	25th	3.29	−0.79	5.47	−12.32	−7.13	−15.96	−10.22
10	50th	3.43	−0.62	6.31	−10.83	−5.70	−13.92	−8.70
	75th	4.19	−0.55	7.04	−9.99	−5.19	−12.90	−8.22
Panel C: Exchange rates (8 series)								
	25th	0.61	−0.53	4.36	−1.73	−1.07	−2.42	−1.60
1	50th	0.62	−0.08	4.51	−1.62	−1.01	−2.10	−1.42
	75th	0.77	0.05	11.60	−1.56	−0.95	−1.92	−1.34
	25th	1.32	−0.36	3.65	−3.72	−2.39	−5.02	−3.36
5	50th	1.39	−0.05	4.05	−3.48	−2.26	−4.33	−3.03
	75th	1.66	0.08	5.91	−3.07	−2.06	−3.82	−2.77
	25th	1.85	−0.31	3.36	−5.00	−3.43	−6.99	−4.55
10	50th	1.93	−0.10	3.52	−4.78	−3.04	−5.72	−4.06
	75th	2.29	0.13	5.12	−4.64	−2.93	−5.41	−3.94

Table 2: Percentage of assets for which the validity of the VaR predictions is rejected
The table presents the percentage of assets for which the unconditional coverage test (UC, Panels A and B) by Kupiec (1995) and the Dynamic Quantile test (DQ, Panels C and D) by Engle and Manganelli (2004) reject the null hypothesis of correct unconditional coverage (UC, DQ) and independence of violations (DQ) for the one–step ahead 1%–VaR (Panels A and C) and 5%–VaR (Panels B and D) at the 5% significance level. The VaR forecasts are obtained for Markov–switching (MS) and single–regime (SR) models for the various universes (426 stocks, 11 indices, and 8 exchange rates) and estimated via Bayesian or frequentist techniques. We highlight in gray the best performing method for the cases in which, for a given asset class and model specification, the percentages of rejections between MS and SR models are significantly different at the 5% level. In the case of stocks, rejections frequencies are corrected for Type I error using the FDR approach of Benjamini and Hochberg (1995).

	Stocks				Stock market indices				Exchange rates			
	Bayesian		Frequentist		Bayesian		Frequentist		Bayesian		Frequentist	
Model	MS	SR	MS	SR	MS	SR	MS	SR	MS	SR	MS	SR
Panel A: UC 1%–VaR												
GARCH $\mathcal{N}$	0.00	26.76	0.23	29.34	72.73	90.91	72.73	90.91	25.00	25.00	25.00	25.00
GARCH sk$\mathcal{N}$	0.00	8.92	0.23	9.62	9.09	63.64	0.00	63.64	0.00	12.50	0.00	12.50
GARCH $\mathcal{S}$	0.00	0.00	0.00	0.00	54.55	45.45	27.27	27.27	25.00	25.00	25.00	12.50
GARCH sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.00	16.43	0.00	19.48	54.55	90.91	63.64	90.91	25.00	25.00	25.00	37.50
GJR sk$\mathcal{N}$	0.00	3.52	0.00	5.16	0.00	54.55	0.00	45.45	0.00	12.50	0.00	25.00
GJR $\mathcal{S}$	0.00	0.00	0.00	0.00	18.18	36.36	18.18	36.36	12.50	12.50	12.50	12.50
GJR sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Panel B: UC 5%–VaR												
GARCH $\mathcal{N}$	0.70	39.20	0.70	38.73	36.36	36.36	27.27	36.36	25.00	50.00	25.00	50.00
GARCH sk$\mathcal{N}$	0.00	41.31	0.00	40.38	0.00	0.00	0.00	0.00	12.50	25.00	0.00	25.00
GARCH $\mathcal{S}$	0.94	1.17	0.70	0.70	54.55	54.55	36.36	54.55	25.00	12.50	25.00	12.50
GARCH sk$\mathcal{S}$	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.47	38.73	0.47	36.15	18.18	18.18	36.36	27.27	25.00	37.50	25.00	37.50
GJR sk$\mathcal{N}$	0.00	40.38	0.00	39.91	0.00	0.00	0.00	0.00	12.50	12.50	0.00	12.50
GJR $\mathcal{S}$	1.64	1.64	0.70	0.47	18.18	27.27	18.18	27.27	37.50	37.50	37.50	37.50
GJR sk$\mathcal{S}$	0.00	0.00	0.00	0.00	0.00	18.18	0.00	18.18	0.00	0.00	0.00	0.00
Panel C: DQ 1%–VaR												
GARCH $\mathcal{N}$	14.08	53.52	14.32	54.69	63.64	90.91	72.73	90.91	25.00	37.50	12.50	37.50
GARCH sk$\mathcal{N}$	14.08	48.36	15.49	50.00	45.45	63.64	45.45	63.64	12.50	37.50	12.50	37.50
GARCH $\mathcal{S}$	19.95	28.64	16.90	29.34	54.55	63.64	63.64	54.55	25.00	25.00	25.00	25.00
GARCH sk$\mathcal{S}$	18.31	23.94	17.37	24.18	45.45	45.45	36.36	36.36	12.50	25.00	12.50	25.00
GJR $\mathcal{N}$	5.87	32.39	6.10	34.74	18.18	90.91	36.36	90.91	12.50	37.50	12.50	37.50
GJR sk$\mathcal{N}$	5.87	27.00	6.10	28.17	9.09	27.27	9.09	45.45	12.50	25.00	0.00	25.00
GJR $\mathcal{S}$	7.04	10.33	4.46	9.86	18.18	27.27	18.18	18.18	12.50	25.00	12.50	25.00
GJR sk$\mathcal{S}$	5.16	10.33	6.57	11.27	0.00	0.00	0.00	0.00	12.50	12.50	12.50	12.50
Panel D: DQ 5%–VaR												
GARCH $\mathcal{N}$	3.52	26.29	3.52	25.82	18.18	9.09	36.36	9.09	0.00	0.00	0.00	0.00
GARCH sk$\mathcal{N}$	3.52	29.81	2.82	30.05	9.09	9.09	9.09	9.09	0.00	0.00	0.00	0.00
GARCH $\mathcal{S}$	1.64	7.75	1.64	8.92	45.45	54.55	36.36	54.55	0.00	0.00	0.00	0.00
GARCH sk$\mathcal{S}$	2.11	6.57	2.82	7.98	9.09	9.09	9.09	9.09	0.00	0.00	0.00	0.00
GJR $\mathcal{N}$	0.00	14.32	0.00	14.55	9.09	9.09	9.09	0.00	0.00	0.00	0.00	0.00
GJR sk$\mathcal{N}$	0.00	15.02	0.00	13.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GJR $\mathcal{S}$	0.00	0.00	0.00	1.17	9.09	0.00	9.09	9.09	12.50	12.50	12.50	12.50
GJR sk$\mathcal{S}$	0.00	0.70	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Standardized gain in average performance when switching from MS to SR models

This table presents the Diebold and Mariano (1995) test statistic of equal average loss function between the MS and SR models for forecasting the distribution of h–day cumulative log–returns ($h \in \{1, 5, 10\}$). As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. Negative values indicate outperformance of the Markov–switching specification compared with the single–regime models. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). The multi–step cumulative log–returns forecasts are generated using 25,000 simulated paths of daily log–returns. Models are estimated with the Bayesian approach.

		Stocks					Stock market indices					Exchange rates				
Horizon	Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
$h = 1$	GARCH $\mathcal{N}$	-0.60	-4.94	-3.74	-8.54	-9.32	-3.84	0.50	-5.11	-1.59	-4.04	-0.09	0.39	-0.90	0.41	-2.65
	GARCH sk$\mathcal{N}$	-0.25	-4.90	-3.43	-8.39	-9.25	-2.64	0.10	-3.32	-1.18	-3.26	0.95	-0.25	-0.77	0.37	-3.41
	GARCH $\mathcal{S}$	-4.00	-3.55	-3.82	-4.42	-3.41	-1.50	-0.90	-1.70	-1.37	-0.17	1.12	-1.26	1.17	1.08	-2.17
	GARCH sk$\mathcal{S}$	-4.52	-4.20	-3.86	-4.86	-2.79	-2.21	-0.87	-1.69	-0.86	0.22	2.18	-0.56	1.63	1.26	-1.45
	GJR $\mathcal{N}$	-0.63	-6.02	-4.07	-10.74	-9.96	-3.58	0.53	-4.99	-2.3	-4.30	0.64	0.58	-0.98	-0.02	-1.64
	GJR sk$\mathcal{N}$	-0.22	-5.95	-3.76	-10.32	-9.94	-2.04	-0.31	-3.06	-1.40	-3.00	0.79	0.07	-0.87	0.11	-1.88
	GJR $\mathcal{S}$	-3.88	-4.44	-3.23	-4.29	-5.00	-1.80	0.49	-2.26	-1.43	0.11	0.36	-1.03	-0.20	-0.19	-2.35
	GJR sk$\mathcal{S}$	-3.64	-4.17	-3.16	-4.10	-3.44	-0.93	0.61	-1.20	-0.55	0.47	0.92	-0.68	1.32	0.65	-1.66
$h = 5$	GARCH $\mathcal{N}$	-0.66	-1.70	-2.16	-7.24	-2.72	-2.11	-1.47	-4.97	-2.48	-1.19	2.69	1.69	-0.52	0.95	0.73
	GARCH sk$\mathcal{N}$	-0.52	-1.67	-1.93	-7.01	-2.65	-2.14	-1.61	-2.69	-1.40	-0.88	2.59	1.17	-0.70	0.43	0.46
	GARCH $\mathcal{S}$	-1.78	-2.27	-2.86	-3.22	-2.68	-0.77	-1.61	-1.52	-2.18	-0.89	2.00	1.53	1.26	1.55	-1.26
	GARCH sk$\mathcal{S}$	-1.70	-2.37	-2.68	-3.23	-2.39	-1.55	-2.72	-0.44	-1.32	-0.32	3.67	0.85	1.02	1.50	-0.93
	GJR $\mathcal{N}$	-0.53	-2.38	-2.24	-8.29	-2.77	0.10	-0.10	-4.91	-3.29	-0.51	3.27	1.15	-1.02	-0.23	0.30
	GJR sk$\mathcal{N}$	-0.30	-2.37	-1.98	-8.37	-2.74	0.62	-1.10	-2.76	-2.44	-0.30	3.70	1.54	-1.09	-0.40	1.15
	GJR $\mathcal{S}$	-1.32	-2.63	-2.46	-2.74	-4.52	-0.21	-1.76	-2.08	-2.85	-0.49	4.08	0.04	0.41	0.75	-2.07
	GJR sk$\mathcal{S}$	-1.14	-1.61	-2.26	-2.75	-3.37	0.26	-0.65	-1.06	-1.40	-0.04	4.60	1.05	1.46	1.11	-1.09
$h = 10$	GARCH $\mathcal{N}$	-0.18	-0.82	-1.59	-6.36	-1.93	-1.23	-0.66	-3.96	-2.07	-1.91	1.55	1.18	-0.33	0.99	0.89
	GARCH sk$\mathcal{N}$	-0.14	-0.71	-1.42	-6.32	-1.96	-1.76	-0.80	-2.69	-2.12	-1.35	1.23	1.58	-0.72	0.25	0.86
	GARCH $\mathcal{S}$	-0.83	-1.01	-1.77	-1.90	-1.25	-1.05	-0.79	-1.69	-2.01	-1.95	1.12	1.63	1.24	1.55	-0.59
	GARCH sk$\mathcal{S}$	-0.93	-1.22	-1.53	-2.04	-1.02	-1.19	-0.99	-1.52	-1.60	-1.22	2.78	2.12	1.30	1.54	-0.46
	GJR $\mathcal{N}$	-0.15	-1.24	-1.64	-6.73	-1.91	0.48	-1.08	-3.87	-2.81	-0.92	1.16	1.11	-0.99	-0.41	0.62
	GJR sk$\mathcal{N}$	0.02	-1.22	-1.32	-6.76	-1.71	0.21	-1.44	-1.79	-1.79	-0.87	2.47	1.15	-0.97	-0.11	1.03
	GJR $\mathcal{S}$	-0.53	-1.46	-1.47	-2.10	-4.04	1.16	-1.43	-1.47	-3.32	-1.13	2.92	0.34	0.71	0.89	-1.71
	GJR sk$\mathcal{S}$	-0.48	-1.15	-1.56	-2.44	-2.80	0.72	-2.19	-1.44	-2.17	-1.31	4.55	1.12	1.79	1.17	-1.09

Table 4: Standardized gain in average performance when switching from MS to SR and changing the specification

This table presents the Diebold and Mariano (1995) test statistic of equal average wCRPS between a MS implementation (in rows) and a SR implementation (in column), for all considered specifications, when forecasting the distribution of one–day ahead log–returns. We report test statistics computed with robust HAC standard errors. Negative values indicate outperformance of the Markov–switching specification compared with single–regime models. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). Models are estimated with the Bayesian approach.

		SR GARCH				SR GJR			
		$\mathcal{N}$	sk$\mathcal{N}$	$\mathcal{S}$	sk$\mathcal{S}$	$\mathcal{N}$	sk$\mathcal{N}$	$\mathcal{S}$	sk$\mathcal{S}$
Panel A: Stocks									
MS GARCH	$\mathcal{N}$	-9.32	-9.56	3.29	3.30	-6.80	-6.85	3.29	3.38
	sk$\mathcal{N}$	-9.00	-9.25	3.60	3.67	-6.60	-6.65	3.42	3.54
	$\mathcal{S}$	-9.01	-9.20	-3.41	-2.99	-7.29	-7.36	-0.14	-0.13
	sk$\mathcal{S}$	-8.86	-9.07	-2.92	-2.79	-7.15	-7.22	0.01	0.04
MS GJR	$\mathcal{N}$	-10.11	-10.26	0.88	0.93	-9.96	-10.25	3.20	3.18
	sk$\mathcal{N}$	-9.88	-10.06	0.88	0.95	-9.64	-9.94	3.33	3.38
	$\mathcal{S}$	-9.73	-9.88	-2.92	-2.76	-9.48	-9.68	-5.00	-4.79
	sk$\mathcal{S}$	-9.57	-9.74	-2.46	-2.34	-9.24	-9.46	-3.19	-3.44
Panel B: Stock market indices									
MS GARCH	$\mathcal{N}$	-4.04	-0.67	3.09	6.00	4.80	7.15	8.15	9.76
	sk$\mathcal{N}$	-5.25	-3.26	-1.04	3.29	3.06	5.46	6.18	8.55
	$\mathcal{S}$	-5.66	-2.90	-0.17	5.09	3.68	6.13	7.17	9.20
	sk$\mathcal{S}$	-6.08	-4.83	-3.52	0.22	2.00	4.39	4.98	7.71
MS GJR	$\mathcal{N}$	-9.65	-7.81	-6.19	-4.26	-4.30	0.33	2.19	4.76
	sk$\mathcal{N}$	-10.39	-9.41	-7.75	-6.35	-5.21	-3.00	-1.80	1.82
	$\mathcal{S}$	-9.79	-8.28	-6.91	-5.11	-4.66	-1.15	0.11	3.92
	sk$\mathcal{S}$	-10.20	-9.53	-8.29	-7.19	-5.34	-3.80	-2.83	0.47
Panel C: Exchange rates									
MS GARCH	$\mathcal{N}$	-2.65	-3.49	5.38	3.95	-2.06	-2.74	3.52	2.81
	sk$\mathcal{N}$	-2.00	-3.41	4.86	5.74	-1.53	-2.45	3.44	3.78
	$\mathcal{S}$	-6.84	-6.53	-2.17	-2.36	-6.09	-6.03	-2.31	-2.45
	sk$\mathcal{S}$	-5.45	-6.29	-0.99	-1.45	-4.81	-5.61	-1.32	-1.73
MS GJR	$\mathcal{N}$	-1.71	-2.33	4.40	3.59	-1.64	-2.35	5.32	3.89
	sk$\mathcal{N}$	-1.13	-1.95	4.26	4.53	-1.02	-1.88	4.53	5.14
	$\mathcal{S}$	-6.02	-6.03	-1.56	-1.68	-6.38	-6.38	-2.35	-2.46
	sk$\mathcal{S}$	-5.05	-5.49	-0.84	-1.21	-5.21	-5.74	-1.35	-1.66

Table 5: Standardized gain in average performance when switching from Bayesian to frequentist estimation
This table presents the Diebold and Mariano (1995) test statistic of equal average loss function between Bayesian and frequentist estimated models for forecasting the distribution of one–day ahead log–returns. As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. Panels A and B report the test statistics when comparing Bayesian against frequentist estimation for SR and MS specifications, respectively. Negative values indicate outperformance of the Bayesian estimation method. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test).

	Stocks					Stock market indices					Exchange rates				
Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
Panel A: Markov–switching GARCH models															
GARCH $\mathcal{N}$	-3.65	-3.26	-2.24	-2.85	-2.24	-0.33	-0.58	0.17	-1.78	-0.25	-1.33	0.99	-1.34	-0.37	-2.02
GARCH sk$\mathcal{N}$	-3.58	-2.93	-2.57	-2.81	-0.60	-1.56	-2.33	-2.05	-2.43	-1.04	-0.82	-1.24	0.97	0.35	-1.04
GARCH $\mathcal{S}$	-2.20	-5.78	-3.12	-4.27	-5.55	0.77	-0.17	-0.89	-0.97	-0.85	-0.78	0.29	-0.69	-1.00	0.35
GARCH sk$\mathcal{S}$	-5.04	-6.88	-6.12	-5.70	-7.04	1.13	-0.52	-0.62	-2.02	-0.58	-1.54	-1.64	-0.35	-1.54	-2.98
GJR $\mathcal{N}$	-1.91	-2.66	-1.59	-2.17	-3.22	-1.21	-2.95	-2.05	-1.81	-2.08	-1.09	-1.38	-0.85	-1.42	-3.61
GJR sk$\mathcal{N}$	-1.83	-3.12	-2.03	-2.44	-2.06	-1.11	-0.84	-2.47	-1.76	-1.40	0.06	-0.32	-0.28	-0.19	-1.17
GJR $\mathcal{S}$	-1.07	-3.11	-2.90	-2.67	-4.48	-1.29	-1.56	-1.66	-2.61	-4.11	-1.75	-2.40	-0.46	-0.51	-4.19
GJR sk$\mathcal{S}$	-3.10	-3.90	-4.67	-2.54	-5.28	-2.95	-2.02	-0.66	-0.46	-3.48	-1.59	-0.38	-0.81	-0.75	-2.50
Panel B: Single–regime GARCH models															
GARCH $\mathcal{N}$	-5.05	-4.23	-6.62	-4.50	-7.84	-2.99	-0.23	-3.40	-3.85	-5.63	-1.59	-0.42	-1.58	0.24	-3.20
GARCH sk$\mathcal{N}$	-4.77	-3.36	-6.59	-4.25	-6.64	-2.55	-1.05	-3.49	-1.98	-4.48	-1.33	-0.86	0.48	-1.00	-4.14
GARCH $\mathcal{S}$	-5.13	-5.08	-5.48	-5.34	-4.93	-1.27	-0.60	-3.01	-1.53	-3.39	-1.41	-1.12	0.15	1.18	-3.76
GARCH sk$\mathcal{S}$	-5.72	-5.40	-5.73	-5.44	-5.18	-2.74	-1.51	0.56	-0.38	-3.87	-2.83	-2.47	-1.74	-1.61	-4.46
GJR $\mathcal{N}$	-5.11	-2.80	-7.64	-3.90	-6.90	-3.65	-2.43	-5.52	-4.87	-5.92	-1.67	-2.70	-1.87	-0.93	-4.14
GJR sk$\mathcal{N}$	-4.55	-2.30	-7.22	-3.02	-5.65	-2.26	-2.03	-2.86	-1.30	-3.94	-0.13	-2.62	-1.53	-0.30	-4.61
GJR $\mathcal{S}$	-3.78	-4.23	-4.90	-4.14	-5.23	-4.13	-2.52	-4.94	-4.00	-4.17	-1.61	-1.96	0.97	-1.29	-4.71
GJR sk$\mathcal{S}$	-3.93	-4.06	-5.32	-4.41	-5.03	-3.82	-1.64	-3.01	-2.01	-3.16	-1.46	-2.24	-0.94	0.74	-4.66

Table 6: Standardized average gain in performance when switching from constrained MS to SR models
This table presents the Diebold and Mariano (1995) test statistic of equal average loss function between the constrained MS and SR models for forecasting the distribution of h–day cumulative log–returns ($h \in \{1, 5, 10\}$). As loss functions, we consider the QL and FZL measures (at $\alpha = 1\%$ and $\alpha = 5\%$), and the wCRPS measure. We report the test statistics computed with robust HAC standard errors, for the time series in the various universes. Negative values indicate outperformance of the shape–parameter constrained Markov–switching specification compared with its single–regime counterpart. In light (dark) gray, we report statistics which are significantly negative (positive) at the 1% level (bilateral test). The multi–step cumulative log–returns forecasts are generated using 25,000 simulated paths of daily log–returns. Models are with the Bayesian approach.

		Stocks					Stock market indices					Exchange rates				
Horizon	Model	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS	QL 1%	QL 5%	FZL 1%	FZL 5%	wCRPS
$h = 1$	GARCH sk$\mathcal{N}$	-0.44	-5.19	-3.56	-8.80	-9.34	-2.68	0.66	-3.53	-1.00	-3.26	0.27	-0.93	-1.13	-0.31	-3.70
	GARCH $\mathcal{S}$	-2.43	-2.32	-3.49	-3.74	-2.97	-1.53	-1.54	-1.68	-1.51	-1.02	0.73	0.10	-0.08	0.25	-0.53
	GARCH sk$\mathcal{S}$	-2.70	-2.69	-3.83	-4.22	-2.62	-1.70	-0.98	-1.33	-0.65	-1.40	1.62	0.31	-0.01	0.54	-0.23
	GJR sk$\mathcal{N}$	-0.37	-6.39	-3.90	-10.91	-9.92	-1.95	-1.59	-3.15	-2.21	-5.07	0.16	-0.91	-1.23	-0.89	-2.77
	GJR $\mathcal{S}$	-2.64	-2.99	-3.29	-3.80	-4.33	-2.25	-0.48	-2.29	-1.44	-0.64	0.30	-0.60	-0.63	-0.22	-1.04
	GJR sk$\mathcal{S}$	-2.90	-3.20	-3.31	-3.73	-4.10	-1.34	-0.93	-1.38	-0.89	-0.71	0.72	-0.40	0.32	-0.17	-0.97
$h = 5$	GARCH sk$\mathcal{N}$	-0.62	-1.79	-1.98	-7.07	-2.76	-2.05	-0.55	-3.02	-1.67	-0.73	2.75	0.63	-1.08	-0.27	0.18
	GARCH $\mathcal{S}$	-0.48	-0.98	-1.74	-1.93	-1.44	-0.92	-2.23	-1.00	-2.43	-1.72	1.96	0.28	-0.58	0.34	-0.58
	GARCH sk$\mathcal{S}$	-0.81	-1.06	-1.98	-2.07	-1.08	-1.58	-2.34	0.16	-1.11	-1.45	1.25	1.56	-0.54	0.65	0.46
	GJR sk$\mathcal{N}$	-0.44	-2.46	-2.11	-8.61	-2.60	-0.48	-1.49	-3.14	-2.55	-0.49	1.94	0.55	-1.26	-1.00	0.41
	GJR $\mathcal{S}$	-0.41	-1.26	-1.94	-2.29	-3.19	-0.19	-1.03	-2.62	-3.49	-0.87	3.19	1.40	-0.61	0.11	-0.56
	GJR sk$\mathcal{S}$	-0.21	-0.83	-1.71	-1.97	-2.86	0.80	-0.33	-1.59	-1.21	0.02	3.02	1.54	0.63	0.24	-0.92
$h = 10$	GARCH sk$\mathcal{N}$	-0.25	-0.84	-1.48	-6.63	-2.06	-1.92	0.03	-3.43	-2.68	-1.71	1.44	1.55	-0.80	-0.09	0.73
	GARCH $\mathcal{S}$	-0.29	-0.25	-0.86	-1.09	-0.23	-1.74	-1.46	-2.00	-2.85	-2.11	0.17	1.76	-0.48	0.40	0.10
	GARCH sk$\mathcal{S}$	-0.27	0.09	-0.61	-0.93	0.42	-0.40	-1.94	-1.01	-2.13	-2.30	1.20	1.30	1.19	1.70	0.92
	GJR sk$\mathcal{N}$	-0.10	-1.28	-1.43	-6.86	-1.80	-0.74	-1.92	-2.28	-2.49	-1.90	0.66	0.61	-1.06	-0.82	0.58
	GJR $\mathcal{S}$	-0.18	-0.41	-0.97	-1.55	-1.74	-0.53	-1.24	-2.23	-3.21	-1.79	2.36	1.27	-0.04	0.24	-0.25
	GJR sk$\mathcal{S}$	0.01	-0.25	-0.91	-1.29	-1.45	0.12	-1.54	-0.95	-1.60	-1.14	2.86	1.06	-0.36	0.47	-1.21

Figure 1: Cumulative performance
This figure presents the evolution of VIX (the Chicago Board of Exchange's volatility index) in the top panel, together with the cumulative loss differentials (QL, FZL and wCRPS) for the 2,000 out–of–sample observations (ranging from December 2008 to November 2016). The comparison is done between the Markov–switching and the single–regime GJR skewed Student–t models. A positive value indicates outperformance of the Markov–switching specification. A positive slope indicates outperformance at the corresponding date.